

On the Consistency of Exact and Approximate Nearest Neighbor with Noisy Data

Wei Gao and Zhi-Hua Zhou*

*National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China*

Abstract

Nearest neighbor has been one of the simplest and most appealing non-parametric approaches in machine learning, pattern recognition, computer vision, etc. Empirical studies have shown the resistance of k -nearest neighbor to noise, yet the theoretical understanding is not clear. This work presents the consistency analysis on exact and approximate nearest neighbor in the random noise setting. Our theoretical studies show that k -nearest neighbor, in the noise setting, gets the same consistent rate as that of noise-free setting, which verifies the robustness of k -nearest neighbor to random noise. The nearest neighbor (1-NN), however, is proven to be biased by random noise. For approximate k -nearest neighbor, we provide a new variant of Johnson-Lindenstrauss lemma, which can be applied to infinite set. Based on this result, we show that the approximate k -nearest neighbor is robust to noise, and achieves better sample complexity, but with a tradeoff between consistency and reduced dimension if there is no additional structural information for the general high-dimensional data. Specifically, approximate k -nearest neighbor with sharp dimensional reduction tends to cause large deviation from the Bayes risk. Finally, we prove the consistency and noisy robustness of approximate k -nearest neighbor for sparse high-dimensional data.

Keywords: Classification, nearest neighbor, random noise, consistency, random projection, Johnson-Lindenstrauss lemma

*Email: zhouzh@lamda.nju.edu.cn

1. Introduction

In many real scenarios, our collected training data are always corrupted by noises, e.g., a document may be mis-classified manually due to human error or bias, a doctor may make incorrect diagnoses for patients because of his knowledge and experience, a spammer can manipulate the data to mislead the outcome of spam-filter systems, etc. Generally speaking, corrupted data may deviate the learning process, increase the sample and model complexities, and deteriorate the quality and effectiveness of learned classifiers; for example, the random noise defeats all convex potential boosters (Long and Servedio, 2010), and support vector machines (SVMs) tend to overfit for noisy labels. The studies on noise have been a valuable topic of great practical importance.

The nearest neighbor (Fix and Hodges, 1951; Cover and Hart, 1967) has been one of the oldest and most intuitive approaches in machine learning, pattern recognition, computer vision, etc. The basic idea is to classify each unlabeled instance by the label of its nearest neighbor (1-NN) or by the majority label of its k nearest neighbors (k -NN) in the training sample. Despite of the simplicity, this approach achieves good performance empirically, makes good explanation for prediction, and has attracted much attention (Wagner, 1971; Kulkarni and Posner, 1995; Dasgupta, 2012; Shalev-Shwartz and Ben-David, 2014; Berlind and Uner, 2015). Various approximate nearest neighbors have been developed to overcome the bottleneck of running time and sample complexity of high-dimensional tasks (Kushilevitz et al., 1998; Ailon and Chazelle, 2006; Har-Peled et al., 2012; Andoni and Razenshteyn, 2015). Empirical studies (Tarlow et al., 2013; Kusner et al., 2014) have shown that k -nearest neighbor tends to be resistant to noise, whereas the theoretical understanding is not clear.

This work studies the binary classification in the presence of label noise, also referred as *random classification noise*. That is, the observed labels have been flipped with some certain probability instead of seeing the true labels. We present the first analysis on the consistency of exact and approximate nearest neighbor in the random noise setting, and our main contributions can be summarized as follows:

- We show that the k -nearest neighbor, in the random noise setting, gets the same consistent rate as that in the noise-free setting, which verifies the robustness of k -nearest neighbor to random noise, especially for large k . The nearest neighbor (1-NN), however, is proven to be

biased by random noise. Our consistency analysis can be easily applied to the noise-free setting, and relevant studies improve the work of (Shalev-Shwartz and Ben-David, 2014).

- We present a variant of Johnson-Lindenstrauss lemma, which can be applied to infinite set. Based on this finding, we analyze the consistency of the proximate k -nearest neighbor without any additional structural information for the general high-dimensional data. We show that the approximate k -nearest neighbor is also robust to random noise as that of the exact k -nearest neighbor, and achieves better sample complexity, but with a tradeoff between consistency and projected dimension. Specifically, approximate k -nearest neighbor with sharp dimension reduction may cause large deviation from Bayes risk.
- For sparse high-dimensional data, we give another variant of Johnson-Lindenstrauss lemma, which is inspired from the restricted isometry property in compressed sensing. Based on this result, we show the consistency of approximate k -nearest neighbor with better sample complexity, and it is also robust to random noise as that of the exact k -nearest neighbor.

The rest of this paper is constructed as follows: Section 2 introduces the relevant work. Section 3 makes some preliminaries. Section 4 provides the consistency analysis of exact nearest neighbor. Sections 5 and 6 provide the consistency analysis of approximate k -nearest neighbor. Section 7 provides detailed proofs, and Section 8 concludes this work.

2. Related work

Angluin and Laird (1988) first proposed the random noise model and proved the PAC-learnable after the pioneer work of PAC learning model (Valiant, 1984). This motivates a series of follow-up theoretical studies on this direction. The finite VC-dimension has been used to characterize the learnability in the work of (Aslam and Decatur, 1996; Cesa-Bianchi et al., 1999) for random noise model. Ben-David et al. (2009) characterized the learnability of online mistake bound model based on the Littlestone dimension. Kearns (1993, 1998) proposed the statistical query (SQ) model by capturing the global statistical properties of large samples rather than individual example. Kalai and Servediob (2005) gave theoretical analysis on boosting algorithms in the presence of random noise.

Various learning algorithms have been developed to deal with noisy data in many real applications, e.g., outlier detection (Brodley and Friedl, 1999), re-weight of training instances (Rebbapragada and Brodley, 2007; Liu and Tao, 2016), perceptron-style algorithms (Bylander, 1994; Crammer et al., 2006; Dredze et al., 2008), robust losses algorithms (Xu et al., 2006; Masnadi-Shirazi and Vasconcelos, 2009; Denchev et al., 2012) as well as unbiased losses methods (Natarajan et al., 2013), etc. The survey articles (Nettleton et al., 2006; Frenay and Verleysen, 2014, reference therein) provided more details on this issue. Empirical studies showed the resistance of k -nearest neighbor to noise (Tarlow et al., 2013; Kusner et al., 2014), whereas the theoretical understanding is not clear.

The study on nearest neighbor could date back to 1950s (Fix and Hodges, 1951), and has attracted much attention (Cover and Hart, 1967; Wagner, 1971; Kulkarni and Posner, 1995; Kpotufe, 2011; Dasgupta, 2012; Dasgupta and Sinha, 2013; Ram and Gray, 2013; Berlind and Uner, 2015). The asymptotic consistency of nearest neighbor has been studied in (Cover and Hart, 1967; Devroye et al., 1994, 1996; Fix and Hodges, 1951; Stone, 1977), and it is well-known that the expected risk converges to the Bayes-optimal risk R^* for the k_n -nearest neighbor if $k_n = o(n)$, to $R^* + O(1/\sqrt{k})$ for the k -nearest neighbor, and to at most $2R^*$ for the nearest neighbor (1-NN). The consistency analysis based on finite sample has also been presented in the work of (Chaudhuri and Dasgupta, 2014; Shalev-Shwartz and Ben-David, 2014). As we know, this still remains open for the consistency of nearest neighbor under the noisy setting.

One drawback of nearest neighbor is that the sample complexity and the requirement of space or query time grow exponentially in the dimensionality both theoretically and empirically (Weber et al., 1998; Har-Peled et al., 2012; Shalev-Shwartz and Ben-David, 2014; Andoni and Razenshteyn, 2015). This phenomenon is also known as the algorithmic “curse of dimensionality”. Many approximate nearest neighbors have been developed to overcome those bottlenecks (Kushilevitz et al., 1998; Ailon and Chazelle, 2006; Andoni and Indyk, 2006; Har-Peled et al., 2012; Andoni and Razenshteyn, 2015), whereas little is known on the consistency of approximate nearest neighbor.

3. Preliminaries

For a real $p \in [0, 1]$, let $\text{Bern}(p)$ denote the Bernoulli distribution with parameter p , and $y \sim \text{Bern}(p)$ represents that the random variable y is drawn according to the Bernoulli distribution with parameter $p \in [0, 1]$. For an

integer $n \geq 0$, let $[n] = \{1, 2, \dots, n\}$, and we denote by $|\mathcal{Z}|$ the cardinality of set \mathcal{Z} . Let $\mathcal{N}(0, 1)$ denote the standard normal distribution. Given a function $f(n)$ and constant $c \in (0, \infty)$, $f(n) = o(n)$, $g(n) = \Omega(n)$, $h(n) = O(n)$ denote $f(n)/n \rightarrow 0$, $g(n)/n \rightarrow \infty$, $h(n)/n \rightarrow c$ as $n \rightarrow \infty$, respectively.

For $p \geq 1$ and vector $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$, the ℓ_p norm is defined as

$$\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \dots + |x_d|^p)^{1/p}.$$

We simplify the Euclidean (ℓ_2) norm as $\|\mathbf{x}\| = \|\mathbf{x}\|_2$, and denote by $\|\mathbf{x}\|_0$ the number of non-zero elements in \mathbf{x} , i.e.,

$$\|\mathbf{x}\|_0 = |\{i \in [d]: x_i \neq 0\}|.$$

Given a matrix A , let A^\top and $\lambda_{\min}(A)$ denote the transpose and minimum eigenvalue of matrix A , respectively. Let \mathbb{I}_d denote the $d \times d$ identity matrix.

Let \mathcal{X} and $\mathcal{Y} = \{+1, -1\}$ denote the input and output space, respectively. Let \mathcal{D} be an (underlying) unknown ground-truth distribution over $\mathcal{X} \times \mathcal{Y}$. Assume that the training data

$$S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$$

are drawn identically and independently (i.i.d.) according to distribution \mathcal{D} . Let $\mathcal{D}_{\mathcal{X}}$ denote the marginal distribution over \mathcal{X} , and let

$$\eta(\mathbf{x}) = \Pr[y = +1 | \mathbf{x}]$$

be the conditional probability with respect to true distribution \mathcal{D} . In this work, we assume that $\eta(\mathbf{x})$ is L -Lipschitz for some constant $L > 0$, that is,

$$|\eta(\mathbf{x}) - \eta(\mathbf{x}')| \leq L \|\mathbf{x} - \mathbf{x}'\|.$$

In this paper, we focus on the ℓ_2 norm, i.e., Euclidean distance, and it is interesting to generalize our analysis to other distance metrics.

Intuitively, this assumption implies that two instances are likely to have similar labels if they are close to each other, and the assumption has been used in binary classification (Cover and Hart, 1967; Shalev-Shwartz and Ben-David, 2014). An interesting future work is to study the consistency of exact and approximate nearest neighbor under some weaker assumptions as in the work of (Chaudhuri and Dasgupta, 2014).

The Bayes classifier and Bayes risk are given, respectively, by

$$h_{\mathcal{D}}^*(\mathbf{x}) = I[\eta(\mathbf{x}) > 1/2] \quad \text{and} \quad R_{\mathcal{D}}^* = E_{\mathbf{x}}[\min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\}].$$

For a hypothesis h , we denote by

$$R_{\mathcal{D}}(h) = E_{(\mathbf{x}, y) \sim \mathcal{D}}[h(\mathbf{x}) \neq y]$$

the expected risk of hypothesis h over the distribution \mathcal{D} .

In the random noise model, each ground-truth label y_i is corrupted independently by a random noise with rate $\rho \in [0, 1/2)$, and we denote \hat{y}_i the corrupted label, i.e.,

$$\hat{y}_i = \begin{cases} -y_i & \text{with probability } \rho, \\ y_i & \text{with probability } 1 - \rho. \end{cases}$$

In this work, we focus on the symmetric noise, that is,

$$\Pr[\hat{y}_i = -1 | y_i = +1] = \Pr[\hat{y}_i = +1 | y_i = -1] = \rho.$$

This model has been well-studied in (Angluin and Laird, 1988).

Let $\hat{\mathcal{D}}$ denote the corrupted distribution. We denote by

$$\hat{S}_n = \{(\mathbf{x}_1, \hat{y}_1), (\mathbf{x}_2, \hat{y}_2), \dots, (\mathbf{x}_n, \hat{y}_n)\}$$

the corrupted sample by random noise. Essentially, each example $(\mathbf{x}_i, \hat{y}_i)$ in \hat{S}_n is drawn i.i.d. according to the corrupted distribution $\hat{\mathcal{D}}$. Let

$$\hat{\eta}(\mathbf{x}) = \Pr[\hat{y} = +1 | \mathbf{x}]$$

denote the conditional probability w.r.t. the corrupted distribution $\hat{\mathcal{D}}$.

4. Consistency of Nearest Neighbor

Given a sample $\hat{S}_n = \{(\mathbf{x}_1, \hat{y}_1), (\mathbf{x}_2, \hat{y}_2), \dots, (\mathbf{x}_n, \hat{y}_n)\}$ and $\mathbf{x} \in \mathcal{X}$, let $\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), \dots, \pi_n(\mathbf{x})$ be a reordering of $\{1, 2, \dots, n\}$ according to the distance of \mathbf{x}_i to \mathbf{x} , i.e.,

$$\|\mathbf{x} - \mathbf{x}_{\pi_i(\mathbf{x})}\| \leq \|\mathbf{x} - \mathbf{x}_{\pi_{i+1}(\mathbf{x})}\| \text{ for } i < n.$$

For k-nearest neighbor algorithm, the output hypothesis $h_{\hat{S}_n}(\mathbf{x})$ is defined as

$$h_{\hat{S}_n}^k(\mathbf{x}) = \begin{cases} +1 & \text{for } \sum_{i=1}^k \hat{y}_{\pi_i(\mathbf{x})} \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

We begin with the consistency analysis of k -nearest neighbour in the random noise setting as follows:

Theorem 1. For $\mathcal{X} = [0, 1]^d$ and $k \geq 8$, let $h_{\hat{S}_n}^k$ be the output hypothesis of applying the k -nearest neighbor (k -NN) algorithm to a corrupted sample $\hat{S}_n = \{(\mathbf{x}_1, \hat{y}_1), (\mathbf{x}_2, \hat{y}_2), \dots, (\mathbf{x}_n, \hat{y}_n)\}$, and assume that the noise rate is ρ . We have

$$\begin{aligned} E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{\hat{S}_n}^k)] &\geq R_{\mathcal{D}}^* \\ E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{\hat{S}_n}^k)] &\leq R_{\mathcal{D}}^* + \frac{2R_{\mathcal{D}}^*}{\sqrt{k}} + \frac{2\rho}{(1-2\rho)\sqrt{k}} + 5 \max\{L, \sqrt{L}\} \sqrt{d} \left(\frac{k}{n}\right)^{\frac{1}{1+d}} \end{aligned}$$

where $R(h_{\hat{S}_n}^k) = E_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{\hat{S}_n}(\mathbf{x}) \neq y]]$, and $R_{\mathcal{D}}^*$ denotes the Bayes's risk with respect to the noise-free distribution \mathcal{D} .

In the random noise setting, Theorem 1 gives

$$\lim_{n \rightarrow \infty} E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{\hat{S}_n}^k)] \leq \begin{cases} R_{\mathcal{D}}^* + O(1/\sqrt{k}) & \text{for } k \geq 8 \\ R_{\mathcal{D}}^* & \text{for } k = o(n) \text{ and } k \rightarrow \infty. \end{cases}$$

It is well-known (Fix and Hodges, 1951; Stone, 1977; Devroye, 1981; Dasgupta, 2012) that, in the noise-free setting, the expected risk of k -nearest neighbor converges to $R_{\mathcal{D}}^*$ for $k = o(n)$, and converges to $R_{\mathcal{D}}^* + O(1/\sqrt{k})$ for constant k . Therefore, we get the same consistent rate for k -nearest neighbor even in the random noise setting, and this theoretical result verifies that k -nearest neighbor is resistant to random noise, especially for large k .

From Theorem 1, we can also derive a new consistent bounds for k -nearest neighbor in the noise-free setting (by setting $\rho = 0$) as follows:

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{\hat{S}_n}^k)] \leq R_{\mathcal{D}}^* + \frac{2R_{\mathcal{D}}^*}{\sqrt{k}} + 5 \max\{L, \sqrt{L}\} \sqrt{d} \left(\frac{k}{n}\right)^{\frac{1}{1+d}} \text{ for } k \geq 8.$$

This result improves the work of (Shalev-Shwartz and Ben-David, 2014, Theorem 19.5), which can be written, with our notations, as

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{\hat{S}_n}^k)] \leq R_{\mathcal{D}}^* + \frac{2\sqrt{2}R_{\mathcal{D}}^*}{\sqrt{k}} + (6L\sqrt{d} + k)n^{-\frac{1}{1+d}} \text{ for } k \geq 10.$$

As can be seen, our work guarantees the asymptotic consistency as $k = o(n)$, while (Shalev-Shwartz and Ben-David, 2014, Theorem 19.5) guarantees this property as $k = o(n^{1/(1+d)})$.

The proof of Theorem 1 relies on a key lemma as follows:

Lemma 1. For $k \geq 8$, let $Z = \sum_{i=1}^k Z_i/k$, where Z_1, Z_2, \dots, Z_k are independent Bernoulli random variables with parameters $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_k$, respectively, i.e., $Z_i \sim \text{Bern}(\hat{p}_i)$ for $i \in [k]$. We set $\hat{p} = \sum_{i=1}^k \hat{p}_i/k$, $p = (\hat{p} - \rho)/(1 - 2\rho)$, and let Bernoulli random variable $y \sim \text{Bern}(p)$. We have

$$\begin{aligned} & E_{Z_1, \dots, Z_k} \Pr_{y \sim \text{Bern}(p)} [y \neq I[Z > 1/2]] \\ & \leq \left(1 + \sqrt{\frac{2}{k}}\right) \Pr_{y \sim \text{Bern}(p)} [y \neq I[\hat{p} > 1/2]] + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)}. \end{aligned}$$

The detailed proofs of Theorem 1 and Lemma 1 are given in Section 7.1.

Now, we consider the consistency of the nearest neighbor (1-NN) in the random noise setting. Given sample $\hat{S}_n = \{(\mathbf{x}_1, \hat{y}_1), (\mathbf{x}_2, \hat{y}_2), \dots, (\mathbf{x}_n, \hat{y}_n)\}$, the output hypothesis $h_{\hat{S}_n}(\mathbf{x})$ is defined as

$$h_{\hat{S}_n}(\mathbf{x}) = \hat{y}_{\pi_1(\mathbf{x})}.$$

We present the consistency analysis of nearest neighbor (1-NN) algorithm in the random noise setting. The detailed proof is given in Section 7.2.

Theorem 2. For $\mathcal{X} = [0, 1]^d$, let $h_{\hat{S}_n}$ be the output hypothesis of applying the nearest neighbour (1-NN) algorithm to a corrupted sample $\hat{S}_n = \{(\mathbf{x}_1, \hat{y}_1), (\mathbf{x}_2, \hat{y}_2), \dots, (\mathbf{x}_n, \hat{y}_n)\}$, and assume that the noise rate is ρ . We have

$$\begin{aligned} E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} [R_{\mathcal{D}}(h_{\hat{S}_n})] & \geq \rho + (1 - 2\rho)R_{\mathcal{D}}^* - \frac{3}{2}(1 - 2\rho)L\sqrt{d}n^{-1/(d+1)} \\ E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} [R_{\mathcal{D}}(h_{\hat{S}_n})] & \leq \rho + 2(1 - 2\rho)R_{\mathcal{D}}^* + \frac{3}{2}(1 - 2\rho)L\sqrt{d}n^{-1/(d+1)} \end{aligned}$$

where $R_{\mathcal{D}}(h_{\hat{S}_n}) = E_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{\hat{S}_n}(\mathbf{x}) \neq y]]$, and $R_{\mathcal{D}}^*$ denotes the Bayes's risk with respect to the noise-free distribution \mathcal{D} .

From this theorem, we have

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} [R_{\mathcal{D}}(h_{\hat{S}_n})] - R_{\mathcal{D}}^* \geq (1 - 2R_{\mathcal{D}}^*)\rho \quad \text{as } n \rightarrow \infty,$$

which shows that the nearest neighbor is deviated by random noise unless the trivial case $R_{\mathcal{D}}^* = 1/2$. Notice that $R_{\mathcal{D}}^* = 1/2$ implies that $\eta(\mathbf{x}) = 1/2$ for each $\mathbf{x} \in \mathcal{X}$, and any classifier $h: \mathcal{X} \rightarrow \{-1, 1\}$ is the Bayes' classifiers. Let

us further consider the specific case $\eta(\mathbf{x})(1 - \eta(x)) = 0$ for each $\mathbf{x} \in \mathcal{X}$. It is easy to get $R_{\mathcal{D}}^* = 0$ in this case, and Theorem 2 gives

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{\hat{S}_n})] \rightarrow \rho \quad \text{as } n \rightarrow \infty,$$

which is obviously biased from the Bayes risk $R_{\mathcal{D}}^* = 0$.

Theorem 2 can be easily applied to noise-free case by setting $\rho = 0$, and we have

$$R_{\mathcal{D}}^* - \frac{3}{2}L\sqrt{dn}^{-1/(d+1)} \leq E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{\hat{S}_n})] \leq 2R_{\mathcal{D}}^* + \frac{3}{2}L\sqrt{dn}^{-1/(d+1)}.$$

Recall that the consistency analysis on nearest neighbor has been studied in (Shalev-Shwartz and Ben-David, 2014, Theorem 19.3), which can be written, with our notations, as

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{\hat{S}_n})] \leq 2R_{\mathcal{D}}^* + 4L\sqrt{dn}^{-1/(d+1)}.$$

As can be seen, our work presents better constant.

5. Consistency of Approximate k -Nearest Neighbor for General High-Dimensional Data

One limitation of nearest neighbor is the exponential sample complexity $\Omega(\exp(d))$ as in Theorems 1 and 2, which is essential from the theory of “no free lunch” (Shalev-Shwartz and Ben-David, 2014). The requirements of space or query time also grow exponentially in the dimension as mentioned in the work of (Har-Peled et al., 2012; Andoni and Razenshteyn, 2015). Those exponential dependence on dimensionality are also known as the algorithmic “curse of dimensionality”. To overcome this challenge, various approximate nearest neighbors have attracted much attention (Ailon and Chazelle, 2006; Andoni and Indyk, 2006; Shakhnarovich et al., 2006; Har-Peled et al., 2012).

We concern the random dimensionality reduction, which can be viewed as a lower representation of high-dimensional data, while preserves the relevant properties on pairwise distances approximately. This method is easy to implement in practice, with lower computational expense in comparison with other reduction methods such as PCA, LDA, etc. It is suitable as a preprocessing step without any requirement on the prior knowledge of data.

The basic idea of random dimensionality reduction is to left multiply a random matrix $A \in \mathbb{R}^{\tau \times d}$ with $\tau \ll d$, where each entry in A is drawn

i.i.d. from some subgaussian distributions such as standard normal and Rademacher distribution. We will present detailed analysis for standard normal distribution, and similar analysis could be made for other distributions.

For random dimensionality reduction, it is natural to recall the original Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984; Dasgupta and Gupta, 2003), which shows that a finite set \mathcal{Z} in Euclidean space can be projected to $O(\epsilon^{-1} \log |\mathcal{Z}|)$ with a distortion of at most $1 + \epsilon$ between pairwise instances, as follows:

Theorem 3. *Let \mathcal{Z} be a finite set. Let $A \in \mathbb{R}^{\tau \times d}$ be a random matrix, whose entries are drawn i.i.d. from distribution $\mathcal{N}(0, 1)$. For any $\epsilon, \delta \in (0, 1)$, the following holds with probability at least $1 - \delta$ over the random choice of A ,*

$$(1 - \epsilon)\tau \|\mathbf{x} - \mathbf{x}'\|_2^2 \leq \|A\mathbf{x} - A\mathbf{x}'\|_2^2 \leq (1 + \epsilon)\tau \|\mathbf{x} - \mathbf{x}'\|_2^2$$

for each $\mathbf{x}, \mathbf{x}' \in \mathcal{Z}$ if $\tau \geq 4 \ln(2|\mathcal{Z}|^2/\delta)/(\epsilon^2 - \epsilon^3)$.

Alon (2003) provided a lower bound to show that the dependence of τ on ϵ and $|\mathcal{Z}|$ is optimal up to some constant. Some variants of Johnson-Lindenstrauss lemma are further presented for ℓ_p norms ($p \in [1, 2]$) in the work of (Ailon and Chazelle, 2006; Matousek, 2008). Most previous studies are restricted to finite set.

For consistency analysis of approximate nearest neighbor, one challenge is how to approximately preserve each pairwise distances for large and infinite sample, because consistency concerns the asymptotic property of approximate nearest neighbor approaching to the optimal Bayes risk in the large sample, or even infinite sample limit. For an infinite set, let us see a negative result as follows:

Lemma 2. *Let $\mathcal{Z} = \{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_1 \leq 1\}$. Let $A \in \mathbb{R}^{\tau \times d}$ be a random matrix, whose entries are drawn i.i.d. from distribution $\mathcal{N}(0, 1)$. For any $p \in [1, 2]$, there exist $\mathbf{x}, \mathbf{x}' \in \mathcal{Z}$ such that*

$$\|A'\mathbf{x} - A'\mathbf{x}'\|_p = 0 \quad \text{yet} \quad \|\mathbf{x} - \mathbf{x}'\|_p \geq 1/\sqrt{d}$$

unless $\tau \geq d$.

The detailed proof is presented in Section 7.3. This lemma shows that it is difficult to project an infinite set into a lower dimension space, as well as keep a distortion of $1 + \epsilon$ between pairwise distances simultaneously.

Now, we introduce a new variant of the Johnson-Lindenstrauss lemma for infinite set, which is crucial to the consistency analysis of approximate k -nearest neighbor, as follows:

Theorem 4. *Let \mathcal{Z} be any subset in \mathbb{R}^d . Let $A \in \mathbb{R}^{\tau \times d}$ be a random matrix, whose entries are drawn i.i.d. from $\mathcal{N}(0, 1)$. For any $\epsilon, \delta \in (0, 1)$, the following holds with probability at least $1 - \delta$ over the random choice of A*

$$\tau \|\mathbf{x} - \mathbf{x}'\|_2^2 - \epsilon \tau \|\mathbf{x} - \mathbf{x}'\|_1^2 \leq \|A\mathbf{x} - A\mathbf{x}'\|_2^2 \leq \tau \|\mathbf{x} - \mathbf{x}'\|_2^2 + \epsilon \tau \|\mathbf{x} - \mathbf{x}'\|_1^2$$

for each $\mathbf{x}, \mathbf{x}' \in \mathcal{Z}$ if $\tau \geq 4 \ln(2d^2/\delta)/(\epsilon^2 - \epsilon^3)$.

We defer the detailed proof to Section 7.4. The technique involves the decomposition of each element into an orthonormal basis and the work of (Indyk and Motwani, 1998). Theorem 4 is irrelevant to the size of set \mathcal{Z} , and can be applied to any infinite set. Based on this finding, we will study the consistency of approximate k -nearest neighbor for general high-dimensional data without any additional structural information.

Given a random matrix A , the reduced and corrupted sample is given by

$$A\hat{S} = \{(A\mathbf{x}_1, \hat{y}_1), (A\mathbf{x}_2, \hat{y}_2), \dots, (A\mathbf{x}_n, \hat{y}_n)\}.$$

For an instance $\mathbf{x} \in \mathcal{X}$, we further denote by $\pi_{A,1}(\mathbf{x}), \pi_{A,2}(\mathbf{x}), \dots, \pi_{A,n}(\mathbf{x})$ a reordering of $\{1, 2, \dots, n\}$ according to the distance of \mathbf{x}_i to \mathbf{x} in the reduced τ -dimensional subspace, i.e.,

$$\|A\mathbf{x} - A\mathbf{x}_{\pi_i(\mathbf{x})}\| \leq \|A\mathbf{x} - A\mathbf{x}_{\pi_{i+1}(\mathbf{x})}\| \quad \text{for } i < n.$$

Generally speaking, $\pi_{A,i}(\mathbf{x})$ is not necessarily the same as $\pi_i(\mathbf{x})$ for $i \in [n]$, yet can be viewed as an approximation with theoretical guarantee. The output hypothesis $h_{A,\hat{S}_n}^k(\mathbf{x})$ of k -nearest neighbor algorithm in the reduced τ -dimensional subspace is defined as

$$h_{A,\hat{S}_n}^k(A\mathbf{x}) = \begin{cases} +1 & \text{for } \sum_{i=1}^k \hat{y}_{\pi_{A,i}(\mathbf{x})} \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

We are interested in the expected risk of approximate k -nearest neighbor

$$R(h_{A,\hat{S}_n}^k) = E_{(\mathbf{x},y) \sim \mathcal{D}} \left[I[h_{A,\hat{S}_n}^k(A\mathbf{x}) \neq y] \right]. \quad (1)$$

Based on Theorem 4, we analyze the consistency of approximate k -nearest neighbor in the reduced τ -dimensional subspace and in the random noise setting as follows:

Theorem 5. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_1 \leq 1\}$ and $A' = A/\sqrt{\tau}$, where each entry in A is drawn i.i.d. from distribution $\mathcal{N}(0, 1)$. For $k \geq 8$, let h_{A', \hat{S}_n}^k be the output hypothesis of applying the k -nearest neighbor to the reduced and corrupted sample $\{(A'\mathbf{x}_1, \hat{y}_1), (A'\mathbf{x}_2, \hat{y}_2), \dots, (A'\mathbf{x}_n, \hat{y}_n)\}$, and assume that the noise rate is ρ . For any $\delta, \epsilon \in (0, 1)$, the following holds with probability at east $1 - \delta$ over the random matrix A

$$\begin{aligned} R_{\mathcal{D}}^* &\leq E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{A', \hat{S}_n}^k)] \leq R_{\mathcal{D}}^* + \frac{2R_{\mathcal{D}}^*}{\sqrt{k}} + 10\sqrt{\tau} \max\{\sqrt{L}, L\} \left(\frac{k}{n}\right)^{\frac{1}{1+\tau}} \\ &\quad + \frac{2\rho}{(1-2\rho)\sqrt{k}} + \epsilon \left(2 + \sqrt{\frac{2}{k}} + 10\sqrt{\tau} \max\{\sqrt{L}, L\} \left(\frac{k}{n}\right)^{\frac{1}{1+\tau}}\right) \end{aligned}$$

if $\tau \geq 4 \ln(2d^2/\delta)/(\epsilon^2 - \epsilon^3)$. Here $R(h_{A', \hat{S}_n}^k)$ is defined in Eqn. (1), and $R_{\mathcal{D}}^*$ denotes the Bayes's risk w.r.t. distribution \mathcal{D} .

The detailed proof is presented in Section 7.5. For approximate k -nearest neighbor, Theorem 5 shows that the noise influence can be illustrated by the term $2\rho/(1-2\rho)\sqrt{k}$, which is the same as that of exact k -nearest neighbor. Thus, approximate k -nearest neighbor is also robust to random noise.

By setting $\rho = 0$, this theorem can be applied to noise-free setting, i.e.,

$$\begin{aligned} E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{A', \hat{S}_n}^k)] &\leq R_{\mathcal{D}}^* + \frac{2R_{\mathcal{D}}^*}{\sqrt{k}} + 10\sqrt{\tau} \max\{\sqrt{L}, L\} \left(\frac{k}{n}\right)^{\frac{1}{1+\tau}} \\ &\quad + \epsilon \left(2 + \sqrt{\frac{2}{k}} + 10\sqrt{\tau} \max\{\sqrt{L}, L\} \left(\frac{k}{n}\right)^{\frac{1}{1+\tau}}\right). \end{aligned}$$

The sample complexity of approximate k -nearest neighbor is $\Omega(\exp(\tau))$, which is tighter than $\Omega(\exp(d))$ of exact nearest neighbor as in Theorem 1. Meanwhile, we should also notice the additional term

$$2\epsilon + \epsilon\sqrt{2/k} + 10\epsilon\sqrt{\tau} \max\{\sqrt{L}, L\} (k/n)^{\frac{1}{1+\tau}}$$

which can be viewed as the cost of using approximate k -nearest neighbor.

For $\tau \geq 4 \ln(2d^2/\delta)/(\epsilon^2 - \epsilon^3)$, Theorem 5 further presents the asymptotic properties of approximate k -nearest neighbor as follows:

$$\lim_{n \rightarrow \infty} E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{\hat{S}_n}^k)] \leq \begin{cases} R_{\mathcal{D}}^* + 2\epsilon + O(1/\sqrt{k}) & \text{for } k \geq 8, \\ R_{\mathcal{D}}^* + 2\epsilon & \text{for } k = o(n) \text{ and } k \rightarrow \infty. \end{cases}$$

As can be seen, approximate k -nearest neighbor has a bias 2ϵ from the Bayes risk, and there is a tradeoff between the bias 2ϵ and the reduced dimension τ . The smaller ϵ , the higher dimension τ , and vice versa. Intuitively, sharp dimension reduction will cause more information loss, and increases the deviation from Bayes risk.

6. Consistency of Approximate k -Nearest Neighbor for Sparse High-Dimensional Data

Fortunately, the high-dimensional data are always accompanied with intrinsic structures in many real applications, such as low-dimension manifold (Roweis and Saul, 2000; Tenenbaum et al., 2000), sparsity (Jing et al., 2007; Dauphin and Bengio, 2013), etc. This section studies the consistency of approximate k -nearest neighbor for sparse high-dimensional data, and an interesting work is to explore other structures in the future.

We consider the s -sparse data for instance space, i.e.,

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_2 \leq 1 \text{ and } \|\mathbf{x}\|_0 \leq s\},$$

where $s \ll d$. How to preserve the distance between each pairwise instances for the sparse data has been well-studied in compressed sensing (Candès and Tao, 2006; Donoho, 2006), which is also known as “restricted isometry property” (Candès, 2008). We introduce a variant of Johnson-Lindenstraus lemma inspired from compressed sensing. The detailed proof is presented in Section 7.6 for completeness.

Theorem 6. *Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_2 \leq 1 \text{ and } \|\mathbf{x}\|_0 \leq s\}$. Let $A \in \mathbb{R}^{\tau \times d}$ be a random matrix, whose entries are drawn i.i.d. from distribution $\mathcal{N}(0, 1)$. For any $\epsilon \in (0, \sqrt{2} - 1)$ and $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$ over the random choice of A ,*

$$\left(1 - \frac{\epsilon}{1 - \sqrt{2}\epsilon}\right)\tau\|\mathbf{x} - \mathbf{x}'\|_2^2 \leq \|A\mathbf{x} - A\mathbf{x}'\|_2^2 \leq \left(1 + \frac{\epsilon}{1 - \sqrt{2}\epsilon}\right)\tau\|\mathbf{x} - \mathbf{x}'\|_2^2$$

for each $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ if $\tau \geq (24s \ln(9d/2s\epsilon) + 12 \ln 4/\delta)/(3\epsilon^2 - 2\epsilon^3)$.

For a sample $A\hat{S} = \{(A\mathbf{x}_1, \hat{y}_1), (A\mathbf{x}_2, \hat{y}_2), \dots, (A\mathbf{x}_n, \hat{y}_n)\}$ and random matrix A , we denote by $\pi_{A,1}(\mathbf{x}), \pi_{A,2}(\mathbf{x}), \dots, \pi_{A,n}(\mathbf{x})$ a reordering of $\{1, 2, \dots, n\}$ according to the distance of \mathbf{x}_i to \mathbf{x} in the reduced τ -dimensional subspace.

The output hypothesis $h_{A, \hat{S}_n}^k(\mathbf{x})$ of k -nearest neighbor algorithm in the reduced τ -dimensional subspace is defined as

$$h_{A, \hat{S}_n}^k(A\mathbf{x}) = \begin{cases} +1 & \text{for } \sum_{i=1}^k \hat{y}_{\pi_{A,i}(\mathbf{x})} \geq 0, \\ -1 & \text{otherwise.} \end{cases}$$

Based on Theorem 6, we analyze the consistency of approximate k -nearest neighbor for sparse high-dimensional data as follows:

Theorem 7. *Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_2 \leq 1 \text{ and } \|\mathbf{x}\|_0 \leq s\}$ and $A' = A/\sqrt{\tau}$, where each entry in A is drawn i.i.d. from distribution $\mathcal{N}(0, 1)$. For $k \geq 8$, let h_{A', \hat{S}_n}^k be the output hypothesis of applying the k -nearest neighbor to the reduced and corrupted sample $\{(A'\mathbf{x}_1, \hat{y}_1), (A'\mathbf{x}_2, \hat{y}_2), \dots, (A'\mathbf{x}_n, \hat{y}_n)\}$, and assume that the noise rate is ρ . For any $\epsilon \in (0, \sqrt{2} - 1)$ and $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$ over the random matrix A*

$$R_{\mathcal{D}}^* \leq E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{A', \hat{S}_n}^k)] \leq R_{\mathcal{D}}^* + 2R_{\mathcal{D}}^*/\sqrt{k} \\ + \frac{2\rho}{(1-2\rho)\sqrt{k}} + 15\sqrt{\tau} \max\{L, \sqrt{L}\} \left(1 + \frac{\epsilon}{1 - \sqrt{2}\epsilon}\right) \left(\frac{k}{n}\right)^{\frac{1}{1+\tau}}$$

if $\tau \geq (24s \ln(9d/2s\epsilon) + 12 \ln 4/\delta)/(3\epsilon^2 - 2\epsilon^3)$. Here $R(h_{A', \hat{S}_n}^k)$ is defined in Eqn. (1), and $R_{\mathcal{D}}^*$ denotes the Bayes's risk w.r.t. distribution \mathcal{D} .

The detailed proof is given in Section 7.7. For sparse high-dimensional data, Theorem 7 shows the noisy robustness of approximate k -nearest neighbor as that of exact k -nearest neighbor. By setting $\rho = 0$, we have

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{A', \hat{S}_n}^k)] \\ \leq R_{\mathcal{D}}^* + \frac{2R_{\mathcal{D}}^*}{\sqrt{k}} + 15\sqrt{\tau} \max\{L, \sqrt{L}\} \left(1 + \frac{\epsilon}{1 - \sqrt{2}\epsilon}\right) \left(\frac{k}{n}\right)^{\frac{1}{1+\tau}}.$$

For sparse data, the sample complexity of approximate k -nearest neighbor is $\Omega(\exp(\tau))$, which is tighter than $\Omega(\exp(d))$ of exact nearest neighbor.

For $\tau \geq (24s \ln(9d/2s\epsilon) + 12 \ln 4/\delta)/(3\epsilon^2 - 2\epsilon^3)$, Theorem 7 further presents the asymptotic properties of approximate k -nearest neighbor as follows:

$$\lim_{n \rightarrow \infty} E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R(h_{\hat{S}_n}^k)] \leq \begin{cases} R_{\mathcal{D}}^* + O(1/\sqrt{k}) & \text{for } k \geq 8, \\ R_{\mathcal{D}}^* & \text{for } k = o(n) \text{ and } k \rightarrow \infty. \end{cases}$$

This shows the asymptotic consistency of approximate k -nearest neighbor for sparse high-dimensional data.

7. Proofs

In this section, we provide detailed proof for our main results.

7.1. Proofs of Theorem 1 and Lemma 1

We begin with the following lemma, which is helpful to the proof of Lemma 1.

Lemma 3. *For $k \geq 8$ and $\hat{p} \in [0, 1/2]$, we have*

$$(1 - 2\hat{p})e^{k(\frac{1}{2}-\hat{p})+\frac{k}{2}\log 2\hat{p}} \leq \sqrt{2}\hat{p}/\sqrt{k}.$$

Proof: We first write

$$f(\hat{p}) = (1 - 2\hat{p})e^{k(\frac{1}{2}-\hat{p})+\frac{k}{2}\log 2\hat{p}}/2\hat{p} = (1 - 2\hat{p})e^{\frac{k}{2}(1-2\hat{p})}(2\hat{p})^{\frac{k}{2}-1}$$

and the derivative is given by

$$f'(\hat{p}) = \frac{2k}{p} \left(\hat{p}^2 - \hat{p} + \frac{1}{4} - \frac{1}{2k} \right) e^{\frac{k}{2}(1-2\hat{p})}(2\hat{p})^{\frac{k}{2}-1}.$$

Solving $f'(\hat{p}) = 0$ gives the optimal solution

$$\hat{p}^* = \frac{1}{2} \left(1 - \sqrt{\frac{2}{k}} \right) \quad \text{for } \hat{p}^* \in [0, 1/2].$$

It is easy to find that

$$f(\hat{p}) \leq \max_{\hat{p} \in [0, 1/2]} f(\hat{p}) = \max\{f(0), f(1/2), f(\hat{p}^*)\} = f(\hat{p}^*) \quad (2)$$

because $f(\hat{p})$ is continuous for $\hat{p} \in [0, 1/2]$. We further have

$$f(\hat{p}^*) = \sqrt{\frac{2}{k}} \left(1 - \sqrt{\frac{2}{k}} \right)^{\frac{k}{2}-1} \exp \left(\frac{\sqrt{2k}}{2} \right) = \sqrt{\frac{2}{k}} \exp(g(k))$$

where

$$g(k) = \sqrt{\frac{2}{k}} + \left(\frac{k}{2} - 1 \right) \ln \left(1 - \sqrt{\frac{2}{k}} \right) \leq 2\sqrt{\frac{2}{k}} - \sqrt{\frac{k}{2}} \leq -1$$

where we use the facts $\ln(1-x) \leq -x$ and $k \geq 8$. Therefore, we have

$$f(\hat{p}^*) \leq \sqrt{2}/e\sqrt{k} \leq \sqrt{2}/\sqrt{k}$$

This lemma follows by combining with Eqn. (2). \square

Proof of Lemma 1 We will present detailed proof for $\hat{p} \leq 1/2$, and similar consideration could be proceeded for $\hat{p} > 1/2$. For $\hat{p} \leq 1/2$, we have

$$\Pr_{y \sim \text{Bern}(p)}[y \neq I[\hat{p} > 1/2]] = p$$

and

$$\begin{aligned} E_{Z_1, \dots, Z_k} \Pr_{y \sim \text{Bern}(p)}[y \neq I[Z > 1/2]] \\ &= p \Pr[Z \leq 1/2] + (1-p) \Pr[Z > 1/2] \\ &= p(1 - \Pr[Z > 1/2]) + (1-p) \Pr[Z > 1/2] \\ &= p + (1-2p) \Pr[Z > 1/2]. \end{aligned}$$

Based on the Chernoff's bound, we have

$$\Pr[Z > 1/2] = \Pr[Z - \hat{p} > 1/2 - \hat{p}] \leq e^{k(\frac{1}{2} - \hat{p}) + \frac{k}{2} \log 2\hat{p}}.$$

For $k \geq 8$, we have

$$\begin{aligned} (1-2p) \Pr[Z > 1/2] &= \frac{1-2\hat{p}}{1-2\rho} \Pr[Z > 1/2] \\ &\leq \frac{1-2\hat{p}}{1-2\rho} e^{k(\frac{1}{2} - \hat{p}) + \frac{k}{2} \log 2\hat{p}} \leq \frac{\sqrt{2}\hat{p}}{(1-2\rho)\sqrt{k}} \end{aligned}$$

where the first equality holds from $1-2p = (1-2\hat{p})/(1-2\rho)$, and the last inequality holds from Lemma 3. We complete the proof from the fact $\hat{p} = p + \rho - 2p\rho$. \square

Before the proof of Theorem 1, we present some useful lemmas. It is obvious to derive the following lemma by simple calculation.

Lemma 4. For $p, \hat{p} \in [0, 1]$ and $\rho \in [0, 1/2)$, let $\hat{p} = p + \rho - 2p\rho$. Then we have

$$p < 1/2 \quad \text{if and only if} \quad \hat{p} < 1/2$$

Lemma 5. For $t \geq 1$, we have

$$(1 + 1/t) t^{\frac{1}{2(t+1)}} \leq 2.$$

Proof: Let $g(t) = (1 + 1/t) t^{\frac{1}{2(t+1)}}$, and we have

$$g'(t) = -t^{\frac{1}{2(t+1)}} \left(\frac{1}{2t^2} + \frac{\ln t}{2t(t+1)} \right) < 0 \text{ for } t \geq 1.$$

Therefore, $g(t)$ is a decreasing function, and $g(t) \leq g(1) = 2$ for $t \geq 1$. This completes the proof as desired. \square

We further introduce two lemmas from (Shalev-Shwartz and Ben-David, 2014) as follows:

Lemma 6. (Shalev-Shwartz and Ben-David, 2014, Lemma 19.6) Denote by C_1, C_2, \dots, C_r a collection of subsets over some domain \mathcal{X} . Let S be a sequence of m samples drawn i.i.d. according to distribution \mathcal{D} over \mathcal{X} . Then, for every $k \geq 2$, we have

$$E_{S \sim \mathcal{D}^m} \left[\sum_{i: |C_i \cap S| < k} P[C_i] \right] \leq \frac{2rk}{m}.$$

Lemma 7. (Shalev-Shwartz and Ben-David, 2014, Exercise 19.3) For $p, \hat{p} \in [0, 1]$ and $y' \in \{0, 1\}$, we have

$$\Pr_{y \sim \text{Bern}(p)} [y \neq y'] \leq \Pr_{y \sim \text{Bern}(p')} [y \neq y'] + |p - p'|.$$

Proof: We have

$$\begin{aligned} & \Pr_{y \sim \text{Bern}(p)} [y \neq y'] - \Pr_{y \sim \text{Bern}(p')} [y \neq y'] \\ &= (p - p')I[y' \neq 1] + (p' - p)I[y' \neq 0] \\ &\leq |p - p'| (I[y' \neq 1] + I[y' \neq 0]) = |p - p'|, \end{aligned}$$

which completes the proof. \square

Proof of Theorem 1 We can easily obtain $E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} [R_{\mathcal{D}}(h_{\hat{S}_n}^k)] \geq R_{\mathcal{D}}^*$ from $R_{\mathcal{D}}(h_{\hat{S}_n}^k) \geq R_{\mathcal{D}}^*$. Fixed $\mu > 0$, and let C_1, \dots, C_r be the cover of instance

space \mathcal{X} using boxes of length μ , where $r = (1/\mu)^d$. For simplicity, we denote by the events

$$\begin{aligned}\Gamma_1(\mathbf{x}, \mathbf{x}') &= \{\text{there exists a } C_i \text{ such that } \mathbf{x} \in C_i \text{ and } \mathbf{x}' \in C_i\}, \\ \Gamma_2(\mathbf{x}, \mathbf{x}') &= \{\text{for any } C_i, \text{ we have either } \mathbf{x} \notin C_i \text{ or } \mathbf{x}' \notin C_i\}.\end{aligned}$$

Based on the total probability theorem, we have

$$\begin{aligned}& \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{\hat{S}_n}^k(\mathbf{x}) \neq y]] \\ &= \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{\hat{S}_n}^k(\mathbf{x}) \neq y] \mid \Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_k(\mathbf{x})})] \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [\Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_k(\mathbf{x})})] \\ &\quad + \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{\hat{S}_n}^k(\mathbf{x}) \neq y] \mid \Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_k(\mathbf{x})})] \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [\Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_k(\mathbf{x})})] \\ &\leq \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{\hat{S}_n}^k(\mathbf{x}) \neq y] \mid \Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_k(\mathbf{x})})] + \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [\Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_k(\mathbf{x})})].\end{aligned}$$

This follows that

$$\begin{aligned}E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} [R_{\mathcal{D}}(h_{\hat{S}_n}^k)] &= E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{\hat{S}_n}^k(\mathbf{x}) \neq y]] \right] \\ &\leq \frac{2rk}{n} + E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{\hat{S}_n}^k(\mathbf{x}) \neq y] \mid \Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_k(\mathbf{x})})] \right],\end{aligned}\quad (3)$$

where the inequality holds from the following fact that, by using Lemma 6,

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [\Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_k(\mathbf{x})})] \right] = E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\sum_{i: C_i \cap \hat{S}_n = \emptyset} P[C_i] \right] \leq \frac{2rk}{n}.$$

To upper bound Eqn.(3), we first fixed the training instances $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ and instance \mathbf{x} , and assume $\mathbf{x}_1, \dots, \mathbf{x}_k$ are the k -nearest neighbors, i.e., $\|\mathbf{x}_i - \mathbf{x}\| \leq \mu\sqrt{d}$ for $i \in [k]$. Let $\eta(\mathbf{x}_1), \dots, \eta(\mathbf{x}_k)$ be the conditional probability w.r.t. distribution \mathcal{D} , and let $\hat{\eta}(\mathbf{x}_1), \dots, \hat{\eta}(\mathbf{x}_k)$ be the conditional probability w.r.t. the corrupted distribution \mathcal{D} . We set $p = \sum_{i=1}^k \eta(\mathbf{x}_i)/k$ and $\hat{p} = \sum_{i=1}^k \hat{\eta}(\mathbf{x}_i)/k$. This follows

$$(1 - 2\rho)p = \hat{p} - \rho \quad (4)$$

since $\hat{\eta}(\mathbf{x}_i) = \eta(\mathbf{x}_i)(1 - \rho) + \rho(1 - \eta(\mathbf{x}_i)) = \eta(\mathbf{x}_i) + \rho - 2\rho\eta(\mathbf{x}_i)$ for each $i \in [k]$. Based on Lemma 7, we have

$$\begin{aligned} & E_{\hat{y}_1 \sim \text{Bern}(\hat{\eta}(\mathbf{x}_1)), \dots, \hat{y}_k \sim \text{Bern}(\hat{\eta}(\mathbf{x}_k))} \left[\Pr_{y \sim \text{Bern}(\eta(\mathbf{x}))} [I[h_{\hat{S}_n}^k(\mathbf{x}) \neq y]] \right] \\ & \leq E_{\hat{y}_1 \sim \text{Bern}(\hat{\eta}(\mathbf{x}_1)), \dots, \hat{y}_k \sim \text{Bern}(\hat{\eta}(\mathbf{x}_k))} \left[\Pr_{y \sim \text{Bern}(p)} [I[h_{\hat{S}_n}^k(\mathbf{x}) \neq y]] \right] + |p - \eta(\mathbf{x})|. \end{aligned} \quad (5)$$

This follows that, from Lemma 1 and Eqn. (4),

$$\begin{aligned} & E_{\hat{y}_1 \sim \text{Bern}(\hat{\eta}(\mathbf{x}_1)), \dots, \hat{y}_k \sim \text{Bern}(\hat{\eta}(\mathbf{x}_k))} \left[\Pr_{y \sim \text{Bern}(p)} [I[h_{\hat{S}_n}^k(\mathbf{x}) \neq y]] \right] \\ & \leq \left(1 + \sqrt{\frac{2}{k}} \right) \Pr_{y \sim \text{Bern}(p)} [y \neq I[\hat{p} > 1/2]] + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)} \\ & = \left(1 + \sqrt{\frac{2}{k}} \right) \Pr_{y \sim \text{Bern}(p)} [y \neq I[p > 1/2]] + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)} \end{aligned} \quad (6)$$

where the last equality holds from Lemma 4. Further, we have

$$\begin{aligned} & \Pr_{y \sim \text{Bern}(p)} [y \neq I[p > 1/2]] = \min\{p, 1 - p\} \\ & \leq \min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\} + |p - \eta(\mathbf{x})|, \end{aligned}$$

which implies, by combining with Eqns. (3), (5) and (6),

$$\begin{aligned} & E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} [R_{\mathcal{D}}(h_{\hat{S}_n}^k)] \\ & \leq \left(1 + \sqrt{\frac{2}{k}} \right) R_{\mathcal{D}}^* + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)} + \frac{2rk}{n} + \left(2 + \sqrt{\frac{2}{k}} \right) |p - \eta(\mathbf{x})| \\ & \leq \left(1 + \sqrt{\frac{2}{k}} \right) R_{\mathcal{D}}^* + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)} + \frac{2k\mu^{-d}}{n} + \left(2 + \sqrt{\frac{2}{k}} \right) L\mu\sqrt{d}. \end{aligned}$$

By setting $\mu = \left(2k\sqrt{d}/(nL(2 + \sqrt{2/k})) \right)^{\frac{1}{1+d}}$, we have

$$\begin{aligned} & E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} [R_{\mathcal{D}}(h_{\hat{S}_n}^k)] \leq \left(1 + \sqrt{\frac{2}{k}} \right) R_{\mathcal{D}}^* + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)} \\ & \quad + \left(2 + \sqrt{\frac{2}{k}} \right) L\sqrt{d} \left(1 + \frac{1}{d} \right) \left(\frac{2\sqrt{d}k}{(2 + \sqrt{2/k})nL} \right)^{\frac{1}{1+d}} \end{aligned}$$

From Lemma 5, we have

$$\begin{aligned}
& \left(2 + \sqrt{\frac{2}{k}}\right) L\sqrt{d} \left(1 + \frac{1}{d}\right) \left(\frac{2\sqrt{d}k}{(2 + \sqrt{2/k})nL}\right)^{\frac{1}{1+d}} \\
& \leq \left(4 + 2\sqrt{\frac{2}{k}}\right) L\sqrt{d} \left(\frac{2k}{(2 + \sqrt{2/k})nL}\right)^{\frac{1}{1+d}} \\
& \leq 5L\sqrt{d} \left(\frac{k}{nL}\right)^{\frac{1}{1+d}} \leq 5 \max\{L, \sqrt{L}\} \sqrt{d} \left(\frac{k}{n}\right)^{\frac{1}{1+d}}
\end{aligned}$$

for $d \geq 1$ and $k \geq 8$. This completes the proof. \square

7.2. Proof of Theorem 2

Before the proof of Theorem 2, we first introduce a lemma as follows:

Lemma 8. *For integer $d \geq 1$, we have*

$$\left(1 + \frac{1}{d}\right) \left(\frac{d}{e}\right)^{\frac{1}{d+1}} \leq \frac{3}{2}$$

Proof: Let $g(d) = (1 + 1/d)(d/e)^{\frac{1}{d+1}}$. Then, we have

$$g'(d) = \frac{1 - \ln d}{d(1 + d)} \left(\frac{d}{e}\right)^{\frac{1}{d+1}}.$$

By setting $g'(d) = 0$, we get $d = e$, and $g(d) \leq g(e) \leq 3/2$. This completes the proof. \square

Proof of Theorem 2. According to $R_{\mathcal{D}}(h_{\hat{S}_n}) = E_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{\hat{S}_n}(\mathbf{x}) \neq y]]$, we observe that $E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{\hat{S}_n})]$ is the probability of training sample $\hat{S}_n \sim \hat{\mathcal{D}}^n$ and $(\mathbf{x}, y) \sim \mathcal{D}$ such that $\hat{y}_{\pi_1(\mathbf{x})}$ is different from y . We have

$$\begin{aligned}
& E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{\hat{S}_n})] = E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[E_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{\hat{S}_n}(\mathbf{x}) \neq y]]] \\
& = E_{\mathbf{x}, \mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{D}_{\mathcal{X}}^{n+1}, y \sim \text{Bern}(\eta(\mathbf{x})), \hat{y} \sim \text{Bern}(\hat{\eta}(\mathbf{x}_{\pi_1(\mathbf{x})})})}[I[\hat{y} \neq y]] \\
& = E_{\mathbf{x}, \mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{D}_{\mathcal{X}}^{n+1}} \left[\Pr_{y \sim \text{Bern}(\eta(\mathbf{x})), \hat{y} \sim \text{Bern}(\hat{\eta}(\mathbf{x}_{\pi_1(\mathbf{x})})})} [I[\hat{y} \neq y]] \right].
\end{aligned}$$

where we should notice that $\hat{y} \sim \hat{\eta}(\mathbf{x}_{\pi_1(\mathbf{x})})$ from the corrupted distribution $\hat{\mathcal{D}}$. Given two instances \mathbf{x} and \mathbf{x}' , we have

$$\begin{aligned}
& \Pr_{y \sim \text{Bern}(\eta(\mathbf{x})), \hat{y}' \sim \text{Bern}(\hat{\eta}(\mathbf{x}'))} [y \neq \hat{y}'] \\
&= \eta(\mathbf{x})(1 - \hat{\eta}(\mathbf{x}')) + \hat{\eta}(\mathbf{x}')(1 - \eta(\mathbf{x})) \\
&= \eta(\mathbf{x}) + \hat{\eta}(\mathbf{x}')(1 - 2\eta(\mathbf{x})) \\
&= \eta(\mathbf{x}) + \eta(\mathbf{x})(1 - 2\eta(\mathbf{x})) + (\hat{\eta}(\mathbf{x}') - \eta(\mathbf{x}))(1 - 2\eta(\mathbf{x})) \\
&= 2\eta(\mathbf{x})(1 - \eta(\mathbf{x})) + (\hat{\eta}(\mathbf{x}') - \eta(\mathbf{x}))(1 - 2\eta(\mathbf{x})).
\end{aligned}$$

For noisy label \hat{y}' , we have

$$\hat{\eta}(\mathbf{x}') = \eta(\mathbf{x}')(1 - \rho) + (1 - \eta(\mathbf{x}'))\rho = \eta(\mathbf{x}')(1 - 2\rho) + \rho,$$

which implies

$$\hat{\eta}(\mathbf{x}') - \eta(\mathbf{x}) = (\eta(\mathbf{x}') - \eta(\mathbf{x}))(1 - 2\rho) + \rho(1 - 2\eta(\mathbf{x})).$$

This follows that

$$\begin{aligned}
& \Pr_{y \sim \text{Bern}(\eta(\mathbf{x})), \hat{y} \sim \text{Bern}(\hat{\eta}(\mathbf{x}_{\pi_1(\mathbf{x})}))} [y \neq \hat{y}] = \rho \\
& + (2 - 4\rho)\eta(\mathbf{x})(1 - \eta(\mathbf{x})) + (\eta(\mathbf{x}_{\pi_1(\mathbf{x})}) - \eta(\mathbf{x}))(1 - 2\eta(\mathbf{x}))(1 - 2\rho).
\end{aligned}$$

Therefore, we have

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} [R_{\mathcal{D}}(h_{\hat{S}_n})] = \rho + (1 - 2\rho)E_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}} [2\eta(\mathbf{x})(1 - \eta(\mathbf{x}))] \quad (7)$$

$$+ E_{\mathbf{x}, \mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{D}_{\mathcal{X}}^{n+1}} [(\eta(\mathbf{x}_{\pi_1(\mathbf{x})}) - \eta(\mathbf{x}))(1 - 2\eta(\mathbf{x}))(1 - 2\rho)]. \quad (8)$$

For Eqn. (7), we have $\eta(\mathbf{x})(1 - \eta(\mathbf{x})) \leq \min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\}$ from $\eta(\mathbf{x}) \in [0, 1]$, and

$$\begin{aligned}
2\eta(\mathbf{x})(1 - \eta(\mathbf{x})) &= 2\min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\}(1 - \min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\}) \\
&= \min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\}(2 - 2\min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\}) \\
&\geq \min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\}
\end{aligned}$$

where the last inequality holds from $\min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\} \leq 1/2$. This follows that

$$R_{\mathcal{D}}^* \leq E_{\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}} [2\eta(\mathbf{x})(1 - \eta(\mathbf{x}))] \leq 2R_{\mathcal{D}}^*. \quad (9)$$

For Eqn. (8), we have

$$\begin{aligned}
& |E_{\mathbf{x}, \mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{D}_{\mathcal{X}}^{n+1}}[(\eta(\mathbf{x}_{\pi_1(\mathbf{x})}) - \eta(\mathbf{x}))(1 - 2\eta(\mathbf{x}))(1 - 2\rho)]| \\
& \leq E_{\mathbf{x}, \mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{D}_{\mathcal{X}}^{n+1}}[(\eta(\mathbf{x}_{\pi_1(\mathbf{x})}) - \eta(\mathbf{x}))(1 - 2\eta(\mathbf{x}))(1 - 2\rho)] \\
& \leq (1 - 2\rho)LE_{\mathbf{x}, \mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{D}_{\mathcal{X}}^{n+1}}[\|\mathbf{x}_{\pi_1(\mathbf{x})} - \mathbf{x}\|] \tag{10} \\
& = (1 - 2\rho)LE_{\mathbf{x}, \hat{S}_n}[\|\mathbf{x}_{\pi_1(\mathbf{x})} - \mathbf{x}\|] \tag{11}
\end{aligned}$$

where the last inequality holds from $|1 - 2\eta(\mathbf{x})| \leq 1$ and the L -Lipschitz assumption of $\eta(\mathbf{x})$. This remains to bound $E_{\mathbf{x}, \hat{S}_n}[\|\mathbf{x}_{\pi_1(\mathbf{x})} - \mathbf{x}\|]$, and we proceed exactly as in (Shalev-Shwartz and Ben-David, 2014). Fixed $\mu > 0$, and let C_1, \dots, C_r be the cover of instance space \mathcal{X} using boxes of length μ , where $r = (1/\mu)^d$. We have $\|\mathbf{x} - \mathbf{x}_{\pi_1(\mathbf{x})}\| \leq \sqrt{d}\mu$ for \mathbf{x} and $\mathbf{x}_{\pi_1(\mathbf{x})}$ in the same box; otherwise, $\|\mathbf{x} - \mathbf{x}_{\pi_1(\mathbf{x})}\| \leq \sqrt{d}$. This follows that

$$\begin{aligned}
& E_{\mathbf{x}, \hat{S}_n} [\|\mathbf{x}_{\pi_1(\mathbf{x})} - \mathbf{x}\|] \\
& \leq E_{\hat{S}_n} \left[\sum_{i=1}^r \Pr[C_i] (\sqrt{d}\mu I[\hat{S}_n \cap C_i \neq \emptyset] + \sqrt{d} I[\hat{S}_n \cap C_i = \emptyset]) \right].
\end{aligned}$$

From the fact that

$$P(C_i)E_{\hat{S}_n}[I[\hat{S}_n \cap C_i = \emptyset]] = P(C_i)(1 - P(C_i))^n \leq 1/ne,$$

we have

$$E_{\mathbf{x}, \hat{S}_n} [\|\mathbf{x}_{\pi_1(\mathbf{x})} - \mathbf{x}\|] \leq \sqrt{d}(\mu + r/ne) = \sqrt{d}(\mu + 1/ne\mu^d)$$

which implies that, by setting $\mu = (d/ne)^{1/(d+1)}$ and from Lemma 8,

$$E_{\mathbf{x}, \hat{S}_n} [\|\mathbf{x}_{\pi_1(\mathbf{x})} - \mathbf{x}\|] \leq \sqrt{d} \left(1 + \frac{1}{d}\right) \left(\frac{d}{ne}\right)^{\frac{1}{d+1}} \leq \frac{3\sqrt{d}}{2n^{\frac{1}{1+d}}}.$$

From Eqn. (11), we have

$$|E_{\mathbf{x}, \mathbf{x}_1, \dots, \mathbf{x}_n \sim \mathcal{D}_{\mathcal{X}}^{n+1}}[(\eta(\mathbf{x}_{\pi_1(\mathbf{x})}) - \eta(\mathbf{x}))(1 - 2\eta(\mathbf{x}))(1 - 2\rho)]| \leq 3\sqrt{d}(1 - 2\rho)L/2n^{\frac{1}{1+d}}.$$

By combining the above with Eqns. (7)-(9), we complete the proof. \square

7.3. Proof of Lemma 2

Let $A \in \mathbb{R}^{\tau \times d}$ denote any random matrix with each entry drawing i.i.d. from Gaussian distribution $\mathcal{N}(0, 1)$. We have

$$[A]^\top A = \sum_{i=1}^{\tau} \mathbf{a}_i^\top \mathbf{a}_i$$

where $\mathbf{a}_1, \dots, \mathbf{a}_\tau$ denote the row vector of matrix A , i.e., $A = [\mathbf{a}_1; \dots; \mathbf{a}_\tau]$. From $\text{rank}(\mathbf{a}_i^\top \mathbf{a}_i) = 1$ and $\text{rank}(B + C) \leq \text{rank}(B) + \text{rank}(C)$, we have

$$\text{rank}([A]^\top A) \leq \tau < d \quad \text{if} \quad \tau < d,$$

which implies that $\lambda_{\min}([A]^\top A) = 0$. Therefore, there is a nonzero vector \mathbf{x}^* such that

$$\|\mathbf{x}^*\|_1 = 1, \quad \text{and} \quad [A\mathbf{x}^*]^\top A\mathbf{x}^* = 0 \quad \text{implying} \quad A\mathbf{x}^* = 0.$$

Denote by $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_d^*)$, and it is obvious that $|x_i| \leq 1$ for $i \in [d]$. We construct two vectors $\mathbf{x} = (x_1, \dots, x_d) \in \mathcal{Z}$ and $\mathbf{x}' = (x'_1, \dots, x'_d) \in \mathcal{Z}$ by

$$x_i = \max(0, x_i^*) \quad \text{and} \quad x'_i = -\min(0, x_i^*) \quad \text{for} \quad i \in [d].$$

This follows $\mathbf{x}^* = \mathbf{x} - \mathbf{x}'$ and $\|A(\mathbf{x} - \mathbf{x}')\| = 0$. For $p \in [1, 2]$, we have

$$\|\mathbf{x} - \mathbf{x}'\|_p \geq \|\mathbf{x} - \mathbf{x}'\|_2 \geq \sqrt{\|\mathbf{x} - \mathbf{x}'\|_1/d} = \sqrt{\|\mathbf{x}^*\|_1/d} = 1/\sqrt{d}$$

which completes the proof. \square

7.4. Proof of Theorem 4

We first introduce a lemma as follows:

Lemma 9. *For every $\epsilon \in (0, 1)$, we have*

$$\frac{1}{2}(1 + \sqrt{1 + 4\epsilon^2})e^{1 - \sqrt{1 + 4\epsilon^2}} \leq e^{-(\epsilon^2 - \epsilon^3)/2}.$$

Proof: Let $g(\epsilon) = (\epsilon^2 - \epsilon^3)/2 + 1 - \sqrt{1 + 4\epsilon^2} + \ln(1 + \sqrt{1 + 4\epsilon^2}) - \ln 2$. We have $g(0) = 0$, and

$$g'(\epsilon) = \epsilon(1 - 3\epsilon/2 - 4/(1 + \sqrt{1 + 4\epsilon^2})).$$

If $\epsilon \in [2/3, 1)$, then we have $g'(\epsilon) < \epsilon(1 - 3\epsilon/2) < 0$; otherwise, we have

$$g'(\epsilon) < \epsilon(1 - 4/(1 + \sqrt{1 + 4\epsilon^2})) \leq \epsilon(1 - 4/(1 + \sqrt{1 + 4(2/3)^2})) = -\epsilon/2 < 0.$$

Thus, we have $g'(\epsilon) < 0$ for $\epsilon \in (0, 1)$, which implies $g(\epsilon) < g(0) = 0$. This lemma follows. \square

Proof of Theorem 4. Let $\mathbf{e}_1 = (1, 0, 0, \dots, 0)$, $\mathbf{e}_2 = (0, 1, 0, \dots, 0)$, \dots , and $\mathbf{e}_d = (0, 0, 0, \dots, 1)$ denote an orthonormal basis of space \mathbb{R}^d . Let $A = (a_{ij})$ be a $\tau \times d$ matrix whose each element is drawn i.i.d. from the standard Gaussian distribution.

For any fixed $j \in [d]$, we have

$$E_A[\|A\mathbf{e}_j\|_2^2] = \sum_{i=1}^{\tau} E_{a_{ij}}[a_{ij}^2] = \tau.$$

For any $\lambda > 0$, it holds that

$$\begin{aligned} \Pr_A[\|A\mathbf{e}_j\|_2^2 \leq (1 - \epsilon)\tau] &= \Pr_A\left[\sum_{i=1}^{\tau} a_{ij}^2 \leq (1 - \epsilon)\tau\right] \\ &= \Pr_A\left[e^{-\lambda \sum_{i=1}^{\tau} a_{ij}^2} \geq e^{-\lambda(1-\epsilon)\tau}\right] \leq e^{\lambda(1-\epsilon)\tau} E\left[e^{-\lambda \sum_{i=1}^{\tau} a_{ij}^2}\right] \end{aligned}$$

where the last inequality holds from the Markov's inequality. Since each a_{ij} is selected i.i.d. from the Gaussian distribution $\mathcal{N}(0, 1)$, we have

$$E\left[e^{-\lambda \sum_{i=1}^{\tau} a_{ij}^2}\right] = \left(E\left[e^{-\lambda a_{1j}^2}\right]\right)^{\tau} = (1 + 2\lambda)^{-\tau/2},$$

where the second equality holds from the fact

$$\begin{aligned} E\left[e^{-\lambda a_{1j}^2}\right] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\lambda t^2} e^{-t^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(\sqrt{2\lambda+1}t)^2/2} dt = \frac{1}{\sqrt{1+2\lambda}}. \end{aligned}$$

By setting $\lambda = \epsilon/2(1 - \epsilon)$, we have

$$\Pr_A[\|A\mathbf{e}_j\|_2^2 \leq (1 - \epsilon)\tau] \leq ((1 - \epsilon)e^{\epsilon})^{\tau/2} \leq e^{-(\epsilon^2 - \epsilon^3)\tau/4}. \quad (12)$$

In a similar manner, we can prove that

$$\Pr_A [\|A\mathbf{e}_j\|_2^2 \geq (1 + \epsilon)\tau] \leq e^{-(\epsilon^2 - \epsilon^3)\tau/4}. \quad (13)$$

For any fixed $j, l \in [d]$ with $j \neq l$, we have

$$E_A[\langle A\mathbf{e}_j, A\mathbf{e}_l \rangle] = \sum_{i=1}^{\tau} E_{a_{ij}, a_{il}}[a_{ij}a_{il}] = 0.$$

For any $\lambda \in (0, 1)$, we have

$$\begin{aligned} & \Pr_A [\langle A\mathbf{e}_j, A\mathbf{e}_l \rangle \leq -\epsilon\tau] \\ &= \Pr_A \left[\sum_{i=1}^{\tau} a_{ij}a_{il} \leq -\epsilon\tau \right] = \Pr_A \left[e^{-\lambda \sum_{i=1}^{\tau} a_{ij}a_{il}} \geq e^{\lambda\epsilon\tau} \right] \\ &\leq e^{-\lambda\epsilon\tau} E \left[e^{-\lambda \sum_{i=1}^{\tau} a_{ij}a_{il}} \right] = e^{-\lambda\epsilon\tau} \left(E \left[e^{-\lambda a_{1j}a_{1l}} \right] \right)^{\tau} \end{aligned}$$

where the inequality and last equality hold from the Markov's inequality and the independence of a_{ij} , respectively. We further have

$$\begin{aligned} E \left[e^{-\lambda a_{1j}a_{1l}} \right] &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\lambda t_1 t_2} e^{-t_1^2/2} e^{-t_2^2/2} dt_1 dt_2 \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-t_2^2/2} e^{\lambda^2 t_2^2/2} \int_{-\infty}^{\infty} e^{-(t_1 + \lambda t_2)^2/2} dt_1 dt_2 \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-\lambda^2)t_2^2/2} dt_2 = \frac{1}{\sqrt{1-\lambda^2}} \end{aligned}$$

By setting $\lambda = (\sqrt{1 + 4\epsilon^2} - 1)/2\epsilon$, we have

$$\begin{aligned} & \Pr_A [\langle A\mathbf{e}_j, A\mathbf{e}_l \rangle \leq -\epsilon\tau] \\ &\leq \left(\frac{1}{2} (1 + \sqrt{1 + 4\epsilon^2}) e^{1 - \sqrt{1 + 4\epsilon^2}} \right)^{\tau/2} \leq e^{-(\epsilon^2 - \epsilon^3)\tau/4} \end{aligned} \quad (14)$$

where the last inequality holds from Lemma 9. In a similar manner, we can prove

$$\Pr_A [\langle A\mathbf{e}_j, A\mathbf{e}_l \rangle \geq \epsilon\tau] \leq e^{-(\epsilon^2 - \epsilon^3)\tau/4}. \quad (15)$$

For any $\mathbf{x} \in \mathcal{Z}$, we can decompose it, according to the orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_d\}$, as

$$\mathbf{x} = \sum_{j=1}^d \beta_j \mathbf{e}_j \quad \text{for } \beta_j \in \mathbb{R}, \quad \text{and} \quad \|\mathbf{x}\|_2^2 = \sum_{j=1}^d \beta_j^2.$$

We further have

$$\begin{aligned} \|A\mathbf{x}\|^2 &= \left\langle \sum_{j=1}^d \beta_j A\mathbf{e}_j, \sum_{l=1}^d \beta_l A\mathbf{e}_l \right\rangle \\ &= \sum_{j=1}^d \beta_j^2 \|A\mathbf{e}_j\|_2^2 + \sum_{j \neq l} 2\beta_j \beta_l \langle A\mathbf{e}_j, A\mathbf{e}_l \rangle. \end{aligned} \quad (16)$$

Based on the union bound, and Eqns. (12) and (13), the following holds with probability at least $1 - 2de^{-(\epsilon^2 - \epsilon^3)\tau/4}$,

$$(1 - \epsilon)\tau\beta_j^2 \leq \beta_j^2 \|A\mathbf{e}_j\|_2^2 \leq (1 + \epsilon)\tau\beta_j^2. \quad (17)$$

For $j \neq l$, we further have, with probability at least $1 - 2(d^2 - d)e^{-(\epsilon^2 - \epsilon^3)\tau/4}$,

$$-|\beta_j \beta_l| \epsilon \tau \leq \beta_j \beta_l \langle A\mathbf{e}_j, A\mathbf{e}_l \rangle \leq |\beta_j \beta_l| \epsilon \tau \quad (18)$$

from Eqns. (14) and (15), and the union bound. Using the union bound again, and substituting Eqns. (17) and (18) into (16) give that, with probability at least $1 - 2d^2e^{-(\epsilon^2 - \epsilon^3)\tau/4}$,

$$\tau \sum_{j=1}^d \beta_j^2 - \tau\epsilon \sum_{j=1}^d \sum_{l=1}^d |\beta_j \beta_l| \leq \|A\mathbf{x}\|_2^2 \leq \tau \sum_{j=1}^d \beta_j^2 + \tau\epsilon \sum_{j=1}^d \sum_{l=1}^d |\beta_j \beta_l|. \quad (19)$$

Recall that $\|\mathbf{x}\|_2^2 = \sum_{j=1}^d \beta_j^2$, and

$$\|\mathbf{x}\|_1^2 = (|\beta_1| + |\beta_2| + \dots + |\beta_d|)^2 = \sum_{j=1}^d \sum_{l=1}^d |\beta_j \beta_l|$$

which completes the proof from Eqn. (19). \square

7.5. Proof of Theorem 5

Let $A \in \mathbb{R}^{\tau \times d}$ be a random matrix with each entry drawing i.i.d. from Gaussian distribution $\mathcal{N}(0, 1)$, and we write $A' = A/\sqrt{\tau}$. For $\epsilon, \delta \in (0, 1)$, we denote by

$$\mathcal{G} = \{A' : \|\mathbf{x} - \mathbf{x}'\|_2^2 - \epsilon\|\mathbf{x} - \mathbf{x}'\|_1^2 \leq \|A'\mathbf{x} - A'\mathbf{x}'\|_2^2 \leq \|\mathbf{x} - \mathbf{x}'\|_2^2 + \epsilon\|\mathbf{x} - \mathbf{x}'\|_1^2\}. \quad (20)$$

Based on Theorem 4, we have $\Pr_{A'}[\mathcal{G}] \leq 1 - \delta$ if $\tau \geq 4 \ln(2d^2/\delta)/(\epsilon^2 - \epsilon^3)$. For any $A' \in \mathcal{G}$, let $\mathcal{X}' = [-\beta, \beta]^\tau$ denote the reduced instance space after random projection from the original space $\mathcal{X} = [0, 1]^d$, and it is easy to get that $\beta = 1 + \epsilon$. Given $\mu > 0$, let C_1, C_2, \dots, C_r denote a cover of space \mathcal{X}' by r disjoint boxes with each side length μ . This follows that $r = (2\beta/\mu)^\tau$.

For any $A' \in \mathcal{G}$, we denote by the events

$$\begin{aligned} \Gamma_1(\mathbf{x}, \mathbf{x}') &= \{\text{there exists a } C_i \text{ such that } A'\mathbf{x} \in C_i \text{ and } A'\mathbf{x}' \in C_i\}, \\ \Gamma_2(\mathbf{x}, \mathbf{x}') &= \{\text{for any } C_i, \text{ we have either } A'\mathbf{x} \notin C_i \text{ or } A'\mathbf{x}' \notin C_i\}. \end{aligned}$$

Based on the total probability theorem, we have

$$\begin{aligned} &\Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y]] \\ &= \Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y] \mid \Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})] \Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[\Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})] \\ &\quad + \Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y] \mid \Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})] \Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[\Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})]. \end{aligned}$$

By using the facts that $\Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y] \mid \Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})] \leq 1$ and $\Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[\Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})] \leq 1$, we have

$$\begin{aligned} \Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y]] &\leq \Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[\Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})] \\ &\quad + \Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y] \mid \Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})]. \end{aligned}$$

Recall that $\mathbf{x}_{\pi_{A', k}(\mathbf{x})}$ denotes the k -nearest neighbor of \mathbf{x} in the reduced τ -dimensional subspace. We have

$$\begin{aligned} E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{A', \hat{S}_n}^k)] &= E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y]] \right] \\ &\leq \frac{2rk}{n} + E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y] \mid \Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})] \right], \quad (21) \end{aligned}$$

where the inequality holds from the following fact that, by using Lemma 6,

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\Pr_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})}) \right] \right] = E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\sum_{i: C_i \cap \hat{S}_n = \emptyset} P[C_i] \right] \leq \frac{2rk}{n}.$$

To upper bound Eqn.(21), we first fixed the training instances $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ and instance \mathbf{x} , and assume $\mathbf{x}_1, \dots, \mathbf{x}_k$ are the k -nearest neighbors in the reduced τ -dimensional space, i.e., $\|A'\mathbf{x}_i - A'\mathbf{x}\| \leq \mu\sqrt{\tau}$ for $i \in [k]$. Let $\eta(\mathbf{x}_1), \dots, \eta(\mathbf{x}_k)$ be the conditional probability w.r.t. distribution \mathcal{D} , and let $\hat{\eta}(\mathbf{x}_1), \dots, \hat{\eta}(\mathbf{x}_k)$ be the conditional probability w.r.t. the corrupted distribution \mathcal{D} . We set $p = \sum_{i=1}^k \eta(\mathbf{x}_i)/k$ and $\hat{p} = \sum_{i=1}^k \hat{\eta}(\mathbf{x}_i)/k$. This follows

$$(1 - 2\rho)p = \hat{p} - \rho \quad (22)$$

since $\hat{\eta}(\mathbf{x}_i) = \eta(\mathbf{x}_i)(1 - \rho) + \rho(1 - \eta(\mathbf{x}_i)) = \eta(\mathbf{x}_i) + \rho - 2\rho\eta(\mathbf{x}_i)$ for each $i \in [k]$. Based on Lemma 7, we have

$$\begin{aligned} & E_{\hat{y}_1 \sim \text{Bern}(\hat{\eta}(\mathbf{x}_1)), \dots, \hat{y}_k \sim \text{Bern}(\hat{\eta}(\mathbf{x}_k))} \left[\Pr_{y \sim \text{Bern}(\eta(\mathbf{x}))} [I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y]] \right] \\ & \leq E_{\hat{y}_1 \sim \text{Bern}(\hat{\eta}(\mathbf{x}_1)), \dots, \hat{y}_k \sim \text{Bern}(\hat{\eta}(\mathbf{x}_k))} \left[\Pr_{y \sim \text{Bern}(p)} [I[h_{A', \hat{S}_n}^k(A'\mathbf{x}) \neq y]] \right] \\ & \quad + |p - \eta(\mathbf{x})|. \end{aligned} \quad (23)$$

This follows that, from Lemma 1 and Eqn. (22),

$$\begin{aligned} & E_{\hat{y}_1 \sim \text{Bern}(\hat{\eta}(\mathbf{x}_1)), \dots, \hat{y}_k \sim \text{Bern}(\hat{\eta}(\mathbf{x}_k))} \left[\Pr_{y \sim \text{Bern}(p)} [I[h_{A', \hat{S}_n}^k(\mathbf{x}) \neq y]] \right] \\ & \leq \left(1 + \sqrt{\frac{2}{k}} \right) \Pr_{y \sim \text{Bern}(p)} [y \neq I[\hat{p} > 1/2]] + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)} \\ & = \left(1 + \sqrt{\frac{2}{k}} \right) \Pr_{y \sim \text{Bern}(p)} [y \neq I[p > 1/2]] + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)} \end{aligned} \quad (24)$$

where the last equality holds from Lemma 4. Further, we have

$$\Pr_{y \sim \text{Bern}(p)} [y \neq I[p > 1/2]] = \min\{p, 1 - p\} \leq \min\{\eta(\mathbf{x}), 1 - \eta(\mathbf{x})\} + |p - \eta(\mathbf{x})|.$$

From $p = \sum_{i=1}^k \eta(\mathbf{x}_i)/k$, we have

$$\begin{aligned}
|p - \eta(\mathbf{x})| &\leq \frac{1}{k} \sum_{i=1}^k |\eta(\mathbf{x}_i) - \eta(\mathbf{x})| \leq \frac{L}{k} \sum_{i=1}^k \|\mathbf{x}_i - \mathbf{x}\| \\
&\leq \frac{L}{k} \sum_{i=1}^k \sqrt{\|A'\mathbf{x}_i - A'\mathbf{x}\|_2^2 + \epsilon \|\mathbf{x}_i - \mathbf{x}\|_1^2} \quad (\text{from Eqn. (20)}) \\
&\leq \frac{L}{k} \sum_{i=1}^k \|A'\mathbf{x}_i - A'\mathbf{x}\|_2 + \epsilon L \leq L(\mu\sqrt{\tau} + \epsilon).
\end{aligned}$$

which implies, by combining with Eqns. (21), (23) and (24),

$$\begin{aligned}
E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{\hat{S}_n}^k)] &\leq \Delta + \frac{2rk}{n} + \left(2 + \sqrt{\frac{2}{k}}\right) |p - \eta(\mathbf{x})| \\
&\leq \Delta + \frac{2k}{n} \left(\frac{2+2\epsilon}{\mu}\right)^\tau + \left(2 + \sqrt{\frac{2}{k}}\right) L(\mu\sqrt{\tau} + \epsilon).
\end{aligned}$$

where

$$\Delta = \left(1 + \sqrt{\frac{2}{k}}\right) R_{\mathcal{D}}^* + \frac{\sqrt{2}\rho}{\sqrt{k}(1-2\rho)}.$$

By setting $\mu = \left(2k\sqrt{\tau}(2+2\epsilon)^\tau / (nL(2 + \sqrt{2/k}))\right)^{\frac{1}{1+\tau}}$, we have

$$E_{\hat{S}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{\hat{S}_n}^k)] \leq \Delta + \left(2 + \sqrt{\frac{2}{k}}\right) \epsilon + \Pi$$

where

$$\begin{aligned}
\Pi &= L\sqrt{\tau}(2+2\epsilon) \left(2 + \sqrt{\frac{2}{k}}\right) \left(1 + \frac{1}{\tau}\right) \left(\frac{k\sqrt{\tau}}{(1+\epsilon)(2 + \sqrt{2/k})nL}\right)^{\frac{1}{1+\tau}} \\
&\leq 10L\sqrt{\tau}(1+\epsilon) \left(\frac{k}{nL}\right)^{\frac{1}{1+\tau}} \leq 10\sqrt{\tau}(1+\epsilon) \max\{L, \sqrt{L}\} \left(\frac{k}{n}\right)^{\frac{1}{1+\tau}}
\end{aligned}$$

Here the first inequality holds from Lemma 5 and $\tau \geq 1$. We complete the proof as desired. \square

7.6. Proof of Theorem 6

Let $\mathcal{X}^1 = \{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_2 \leq 1 \text{ and } \|\mathbf{x}\|_0 \leq 2s\}$. It suffices to prove that, with probability at least $1 - 4(9d/2s\epsilon)^{2s}e^{-\tau(\epsilon^2/4 - \epsilon^3/6)}$,

$$\sup_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathcal{X}^1} |\mathbf{x}^\top A^\top A \mathbf{x} - \tau| \leq \frac{\tau\epsilon}{1 - \sqrt{2}\epsilon}$$

where \mathbb{I}_d denotes the identity matrix of size $d \times d$.

We first observe

$$\begin{aligned} \sup_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathcal{X}^1} |\mathbf{x}^\top A^\top A \mathbf{x} - \tau| &= \sup_{\|\mathbf{x}\|=1, \mathbf{x} \in \mathcal{X}^1} |\mathbf{x}^\top (A^\top A - \tau \mathbb{I}_d) \mathbf{x}| \\ &\leq \sup_{\mathbf{x} \in \mathcal{X}^1} |\mathbf{x}^\top (A^\top A - \tau \mathbb{I}_d) \mathbf{x}| \leq \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}^1} \mathbf{x}^\top (A^\top A - \tau \mathbb{I}_d) \mathbf{x}' \\ &= \sup_{\mathbf{x} \in \mathcal{X}^1} \sup_{\mathbf{x}' \in \mathcal{X}^1} \mathbf{x}^\top (A^\top A - \tau \mathbb{I}_d) \mathbf{x}'. \end{aligned}$$

For any fixed $\mathbf{x} \in \mathcal{X}^1$, let

$$G(\mathbf{x}) = \sup_{\mathbf{x}' \in \mathcal{X}^1} \mathbf{x}^\top (A^\top A - \tau \mathbb{I}_d) \mathbf{x}'.$$

Denote by \mathcal{X}_ϵ^1 a proper ϵ -net of \mathcal{X}^1 with the smallest cardinality. Then, the cover number $\mathcal{C}(\mathcal{X}^1)$ over \mathcal{X}^1 satisfies

$$\mathcal{C}(\mathcal{X}^1) \leq |\mathcal{X}_\epsilon^1| \leq \left(\frac{9d}{2s\epsilon} \right)^{2s}$$

from the work of (Plan and Vershynin, 2011, Lemma 3.3). We further set

$$G_\epsilon(\mathbf{x}) = \sup_{\mathbf{x}' \in \mathcal{X}_\epsilon^1} \mathbf{x}^\top (A^\top A - \mathbb{I}_d) \mathbf{x}'.$$

From the work of (Koltchinskii, 2011, Lemma 9.2), it is easy to find the relationship between $G(\mathbf{x})$ and $G_\epsilon(\mathbf{x})$ as follows

$$G(\mathbf{x}) \leq G_\epsilon(\mathbf{x}) / (1 - \sqrt{2}\epsilon).$$

By using the union bound and Lemma 10, we have

$$G_\epsilon(\mathbf{x}) \leq \tau\epsilon$$

with probability at least $1 - 4(9d/2s\epsilon)^{2s}e^{-\tau(\epsilon^2/4 - \epsilon^3/6)}$. □

Lemma 10. *For any fixed $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^1$, we have*

$$\Pr_A [|\mathbf{x}^\top A^\top A \mathbf{x}' - \tau \mathbf{x}^\top \mathbb{I}_d \mathbf{x}'| \geq \tau \epsilon] \leq 4e^{-\tau(\epsilon^2/4 - \epsilon^3/6)}.$$

Proof: From the proof of (Indyk and Motwani, 1998; Dasgupta and Gupta, 2003), it is easy to get that

$$\begin{aligned} (1 - \epsilon)\tau \|\mathbf{x} + \mathbf{x}'\|_2^2 &\leq \|A(\mathbf{x} + \mathbf{x}')\|_2^2 \leq (1 + \epsilon)\tau \|\mathbf{x} + \mathbf{x}'\|_2^2 \\ (1 - \epsilon)\tau \|\mathbf{x} - \mathbf{x}'\|_2^2 &\leq \|A(\mathbf{x} - \mathbf{x}')\|_2^2 \leq (1 + \epsilon)\tau \|\mathbf{x} - \mathbf{x}'\|_2^2 \end{aligned}$$

with probability at least $1 - 4e^{-\tau(\epsilon^2/4 - \epsilon^3/6)}$. Then, we have

$$\begin{aligned} \mathbf{x}^\top A^\top A \mathbf{x}' &= (\|A(\mathbf{x} + \mathbf{x}')\|_2^2 - \|A(\mathbf{x} - \mathbf{x}')\|_2^2)/4 \\ &\leq ((1 + \epsilon)\tau \|\mathbf{x} + \mathbf{x}'\|_2^2 - \tau(1 - \epsilon)\|\mathbf{x} - \mathbf{x}'\|_2^2)/4 \\ &\leq \tau \mathbf{x}^\top \mathbf{x}' - \epsilon\tau(\|\mathbf{x}\|_2^2 + \|\mathbf{x}'\|_2^2)/2 \\ &\leq \tau \mathbf{x}^\top \mathbf{x}' - \tau\epsilon|\mathbf{x}^\top \mathbf{x}'| \end{aligned}$$

and prove $\mathbf{x}^\top A^\top A \mathbf{x}' \geq \tau \mathbf{x}^\top \mathbf{x}' + \tau\epsilon|\mathbf{x}^\top \mathbf{x}'|$ similarly. This follows that

$$|\mathbf{x}^\top A^\top A \mathbf{x}' - \tau \mathbf{x}^\top \mathbf{x}'| \leq \tau\epsilon|\mathbf{x}^\top \mathbf{x}'| \leq \tau\epsilon \text{ for } \|\mathbf{x}\| \leq 1 \text{ and } \|\mathbf{x}'\| \leq 1$$

with probability at least $1 - 4e^{-\tau(\epsilon^2/4 - \epsilon^3/6)}$. This completes the proof. \square

7.7. Proof of Theorem 7

This proof is similar to that of Theorem 5. Let $A' = A/\sqrt{\tau}$, where each entry in A is drawn i.i.d. from Gaussian distribution $\mathcal{N}(0, 1)$. For $0 < \epsilon < \sqrt{2} - 1$, let

$$\begin{aligned} \mathcal{G} &= \{A' : \|\mathbf{x} - \mathbf{x}'\|_2^2 - \epsilon\|\mathbf{x} - \mathbf{x}'\|_2^2/(1 - \sqrt{2}\epsilon) \\ &\leq \|A'\mathbf{x} - A'\mathbf{x}'\|_2^2 \leq \|\mathbf{x} - \mathbf{x}'\|_2^2 + \epsilon\|\mathbf{x} - \mathbf{x}'\|_2^2/(1 - \sqrt{2}\epsilon)\}. \end{aligned} \quad (25)$$

We have $\Pr_{A'}[\mathcal{G}] \leq 1 - \delta$ for $\tau \geq (24s \ln(9d/2s\epsilon) + 12 \ln 4/\delta)/(3\epsilon^2 - 2\epsilon^3)$ from Theorem 6. For any $A' \in \mathcal{G}$, let $\mathcal{X}' = \{\mathbf{x} : \|\mathbf{x}\| \leq \beta\}$ denote the reduced instance space after random projection of original space \mathcal{X} . It is easy to get $\beta = 1 + \epsilon/(1 - \sqrt{2}\epsilon)$. Given $\mu > 0$, let C_1, C_2, \dots, C_r denote a cover of space \mathcal{X}' by r disjoint boxes with each side length μ . This follows $r = (3\beta/\mu)^\tau$.

For any $A' \in \mathcal{G}$, we denote by the events

$$\begin{aligned} \Gamma_1(\mathbf{x}, \mathbf{x}') &= \{\text{there exists a } C_i \text{ such that } A'\mathbf{x} \in C_i \text{ and } A'\mathbf{x}' \in C_i\}, \\ \Gamma_2(\mathbf{x}, \mathbf{x}') &= \{\text{for any } C_i, \text{ we have either } A'\mathbf{x} \notin C_i \text{ or } A'\mathbf{x}' \notin C_i\}. \end{aligned}$$

Based on the total probability theorem, we have

$$\begin{aligned} \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{A', \hat{S}_n}^k(A' \mathbf{x}) \neq y]] &\leq \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [\Gamma_2(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})] \\ &\quad + \Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{A', \hat{S}_n}^k(A' \mathbf{x}) \neq y] \mid \Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})]. \end{aligned}$$

This follows that, from Lemma 6,

$$\begin{aligned} E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} [R_{\mathcal{D}}(h_{A', \hat{S}_n}^k)] &= E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{A', \hat{S}_n}^k(A' \mathbf{x}) \neq y]] \right] \\ &\leq \frac{2rk}{n} + E_{\hat{S}_n \sim \hat{\mathcal{D}}^n} \left[\Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [I[h_{A', \hat{S}_n}^k(A' \mathbf{x}) \neq y] \mid \Gamma_1(\mathbf{x}, \mathbf{x}_{\pi_{A', k}(\mathbf{x})})] \right], \quad (26) \end{aligned}$$

To upper bound Eqn.(26), we fix the instances $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ and \mathbf{x} . Without loss of generality, we assume $\mathbf{x}_1, \dots, \mathbf{x}_k$ are the k -nearest neighbors in the reduced τ -dimensional space, i.e., $\|A' \mathbf{x}_i - A' \mathbf{x}\| \leq \mu \sqrt{\tau}$ for $i \in [k]$. Let $\eta(\mathbf{x}_1), \dots, \eta(\mathbf{x}_k)$ be the conditional probability w.r.t. distribution \mathcal{D} , and let $\hat{\eta}(\mathbf{x}_1), \dots, \hat{\eta}(\mathbf{x}_k)$ be the conditional probability w.r.t. the corrupted distribution \mathcal{D} . We set $p = \sum_{i=1}^k \eta(\mathbf{x}_i)/k$ and $\hat{p} = \sum_{i=1}^k \hat{\eta}(\mathbf{x}_i)/k$. This follows

$$(1 - 2\rho)p = \hat{p} - \rho \quad (27)$$

Based on Lemma 7, we have

$$\begin{aligned} E_{\hat{y}_1 \sim \text{Bern}(\hat{\eta}(\mathbf{x}_1)), \dots, \hat{y}_k \sim \text{Bern}(\hat{\eta}(\mathbf{x}_k))} &\left[\Pr_{y \sim \text{Bern}(\eta(\mathbf{x}))} [I[h_{A', \hat{S}_n}^k(A' \mathbf{x}) \neq y]] \right] \\ &\leq E_{\hat{y}_1 \sim \text{Bern}(\hat{\eta}(\mathbf{x}_1)), \dots, \hat{y}_k \sim \text{Bern}(\hat{\eta}(\mathbf{x}_k))} \left[\Pr_{y \sim \text{Bern}(p)} [I[h_{A', \hat{S}_n}^k(A' \mathbf{x}) \neq y]] \right] \\ &\quad + |p - \eta(\mathbf{x})|. \quad (28) \end{aligned}$$

This follows that, from Lemma 1 and Eqn. (27),

$$\begin{aligned} E_{\hat{y}_1 \sim \text{Bern}(\hat{\eta}(\mathbf{x}_1)), \dots, \hat{y}_k \sim \text{Bern}(\hat{\eta}(\mathbf{x}_k))} &\left[\Pr_{y \sim \text{Bern}(p)} [I[h_{A', \hat{S}_n}^k(\mathbf{x}) \neq y]] \right] \\ &\leq \left(1 + \sqrt{\frac{2}{k}} \right) \Pr_{y \sim \text{Bern}(p)} [y \neq I[\hat{p} > 1/2]] + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)} \\ &= \left(1 + \sqrt{\frac{2}{k}} \right) \Pr_{y \sim \text{Bern}(p)} [y \neq I[p > 1/2]] + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)} \quad (29) \end{aligned}$$

where the last equality holds from Lemma 4. Further, we have

$$\Pr_{y \sim \text{Bern}(p)}[y \neq I[p > 1/2]] = \min\{p, 1-p\} \leq \min\{\eta(\mathbf{x}), 1-\eta(\mathbf{x})\} + |p - \eta(\mathbf{x})|.$$

From $p = \sum_{i=1}^k \eta(\mathbf{x}_i)/k$, we have

$$\begin{aligned} |p - \eta(\mathbf{x})| &\leq \frac{1}{k} \sum_{i=1}^k |\eta(\mathbf{x}_i) - \eta(\mathbf{x})| \leq \frac{L}{k} \sum_{i=1}^k \|\mathbf{x}_i - \mathbf{x}\| \\ &\leq \frac{L}{k} \sum_{i=1}^k \sqrt{\frac{1 - \sqrt{2}\epsilon}{1 - (\sqrt{2} + 1)\epsilon}} \|A' \mathbf{x}_i - A' \mathbf{x}\|_2 \quad (\text{from Eqn. (25)}) \\ &\leq L\mu\sqrt{\tau} \sqrt{\frac{1 - \sqrt{2}\epsilon}{1 - (\sqrt{2} + 1)\epsilon}}. \end{aligned}$$

which implies, by combining with Eqns. (26), (28) and (29),

$$\begin{aligned} E_{\hat{\mathcal{S}}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{\hat{\mathcal{S}}_n}^k)] &\leq \Delta + \frac{2rk}{n} + \left(2 + \sqrt{\frac{2}{k}}\right) |p - \eta(\mathbf{x})| \\ &\leq \Delta + \frac{2k}{n} \left(\frac{3\beta}{\mu}\right)^\tau + L\mu\sqrt{\tau} \left(2 + \sqrt{\frac{2}{k}}\right). \end{aligned}$$

where $\beta = 1 + \epsilon/(1 - \sqrt{2}\epsilon)$ and

$$\Delta = \left(1 + \sqrt{\frac{2}{k}}\right) R_{\mathcal{D}}^* + \frac{\sqrt{2}\rho}{\sqrt{k}(1 - 2\rho)}.$$

By setting $\mu = \left(2k\sqrt{\tau}(3\beta)^\tau/(nL(2 + \sqrt{2/k}))\right)^{\frac{1}{1+\tau}}$, we have

$$E_{\hat{\mathcal{S}}_n \sim \hat{\mathcal{D}}^n}[R_{\mathcal{D}}(h_{\hat{\mathcal{S}}_n}^k)] \leq \Delta + \left(2 + \sqrt{\frac{2}{k}}\right) \epsilon + \Pi$$

where

$$\begin{aligned} \Pi &= 3\beta L\sqrt{\tau} \left(2 + \sqrt{\frac{2}{k}}\right) \left(1 + \frac{1}{\tau}\right) \left(\frac{2k\sqrt{\tau}}{3\beta(2 + \sqrt{2/k})nL}\right)^{\frac{1}{1+\tau}} \\ &\leq 15L\beta\sqrt{\tau} \left(\frac{k}{nL}\right)^{\frac{1}{1+\tau}} \leq 15\sqrt{\tau} \left(1 + \frac{\epsilon}{1 - \sqrt{2}\epsilon}\right) \max\{L, \sqrt{L}\} \left(\frac{k}{n}\right)^{\frac{1}{1+\tau}} \end{aligned}$$

Here the first inequality holds from Lemma 5, $\tau \geq 1$ and $k \geq 8$. We complete the proof as desired. \square

8. Conclusion

The nearest neighbor has been one of the oldest, simplest and most intuitive approaches in machine learning, pattern recognition, computer vision, etc. Empirical studies shows that k -nearest neighbor is robust to noise, yet the theoretical understanding is not clear. This work presents the first consistency analysis of exact and approximate nearest neighbor with noisy data. Our theoretical studies show that k -nearest neighbor, in the noise setting, gets the same consistent rate as that in the noise-free setting, which theoretically verifies the robustness of k -nearest neighbor to random noise. The nearest neighbor, however, is proven to be biased by random noise. For approximate k -nearest neighbor, we provide a new variant of Johnson-Lindenstrauss lemma for infinite set. Based on this result, we show that the approximate k -nearest neighbor is robust to noise, and achieves better sample complexity, but with a tradeoff between consistency and reduced dimension if there is no additional structural information for the general high-dimensional data. Specifically, approximate k -nearest neighbor with sharp dimensional reduction tends to cause large deviation from the Bayes risk. Finally, we prove the consistency and noisy robustness of approximate k -nearest neighbor for sparse high-dimensional data. An interesting future work is to study the consistency of exact and approximate nearest neighbor under other noise settings such as the white and Gaussian noise.

References

- Ailon, N. and Chazelle, B. (2006). Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 557–563, Seattle, WA.
- Alon, N. (2003). Problems and results in extremal combinatorics. *Discrete Mathematics*, 273(1):31–53.
- Andoni, A. and Indyk, P. (2006). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 459–468, Berkeley, CA.

- Andoni, A. and Razenshteyn, I. (2015). Optimal data-dependent hashing for approximate near neighbors. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pages 793–801, Portland, OR.
- Angluin, D. and Laird, P. (1988). Learning from noisy examples. *Machine Learning*, 4(2):343–370.
- Aslam, J. and Decatur, S. (1996). On the sample complexity of noise-tolerant learning. *Information Processing Letters*, 57(4):189–195.
- Ben-David, S., Pál, D., and Shalev-Shwartz, S. (2009). Agnostic online learning. In *Proceedings of the 22nd Conference on Learning Theory*, Montreal, Canada.
- Berlind, C. and Uner, R. (2015). Active nearest neighbors in changing environments. In *Proceeding of the the 32nd International Conference on Machine Learning*, pages 1870–1879, Lille, France.
- Brodley, C. and Friedl, M. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167.
- Bylander, T. (1994). Learning linear threshold functions in the presence of classification noise. In *Proceeding of the 7th Conference on Learning Theory*, New York, NY.
- Candès, E. (2008). The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathématique*, 346(9-10):589–592.
- Candès, E. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425.
- Cesa-Bianchi, N., Dichterman, E., Fischer, P., Shamir, E., and Simon, H. (1999). Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM*, 46(5):684–719.
- Chaudhuri, K. and Dasgupta, S. (2014). Rates of convergence for nearest neighbor classification. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 27*, pages 3437–3445. MIT Press, Cambridge, MA.

- Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.
- Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.
- Dasgupta, S. (2012). Consistency of nearest neighbor classification under selective sampling. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 18.1–18.15, Edinburgh, Scotland.
- Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65.
- Dasgupta, S. and Sinha, K. (2013). Randomized partition trees for exact nearest neighbor search. In *Proceeding of the 26th Conference on Learning Theory*, Princeton, NJ.
- Dauphin, Y. and Bengio, Y. (2013). Stochastic ratio matching of rbms for sparse high-dimensional inputs. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 26*, pages 1340–1348. MIT Press, Cambridge, MA.
- Denchov, V., Ding, N., Neven, H., and Vishwanathan, S. (2012). Robust classification with adiabatic quantum optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 863–870, Edinburgh, Scotland.
- Devroye, L. (1981). On the inequality of cover and hart in nearest neighbor discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(1):75–78.
- Devroye, L., Györfi, L., Krzyżak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385.
- Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

- Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306.
- Dredze, M., Crammer, K., and Pereira, F. (2008). Confidence-weighted linear classification. In *Proceedings of the 25th International Conference on Machine Learning*, pages 264–271, Helsinki, Finland.
- Fix, E. and Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination. *USAF School of Aviation Medicine, Randolph Field, Texas. Project 21-49-004, Report 4, Contract AD41(128)-31.*
- Frenay, B. and Verleysen, M. (2014). Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869.
- Har-Peled, S., Indyk, P., and Motwani, R. (2012). Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350.
- Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the 30th annual ACM symposium on Theory of computing*, pages 604–613, Dallas, TX.
- Jing, L., Ng, M., and Huang, J. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, 19(8):1026–1041.
- Johnson, W. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26:189–206.
- Kalai, A. and Servediob, R. (2005). Boosting in the presence of noise. *Journal of Computer and System Sciences*, 71:266–290.
- Kearns, M. (1993). Efficient noise-tolerant learning from statistical queries. In *Proceedings of the 25th Annual ACM Symposium on Theory of Computing*, pages 392–401, San Diego, CA.
- Kearns, M. (1998). Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006.
- Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Springer, Verlag.

- Kpotufe, S. (2011). k -nn regression adapts to local intrinsic dimension. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 19*, pages 729–737. MIT Press, Cambridge, MA.
- Kulkarni, S. and Posner, S. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039.
- Kushilevitz, E., Ostrovsky, R., and Rabani, Y. (1998). Efficient search for approximate nearest neighbor in high dimensional spaces. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, pages 614–623, Dallas, TX.
- Kusner, M., Tyree, S., Weinberger, K., and Agrawal, K. (2014). Stochastic neighbor compression. In *Proceedings of the 31st International Conference on Machine Learning*, pages 622–630, Beijing, China.
- Liu, T. and Tao, D. (2016). Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461.
- Long, P. and Servedio, R. (2010). Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304.
- Masnadi-Shirazi, H. and Vasconcelos, N. (2009). On the design of loss functions for classification: Theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems 22*, pages 1049–1056. MIT Press, Cambridge, MA.
- Matousek, J. (2008). On variants of the johnsonclindenstrauss lemma. *Random Structures and Algorithms*, 33(2):142–156.
- Natarajan, N., Dhillon, I., Ravikumar, P., and Tewari, A. (2013). Learning with noisy labels. In *Advances in Neural Information Processing Systems 26*, pages 1196–1204. MIT Press, Cambridge, MA.
- Nettleton, D., Orriolspuig, A., and Fornells, A. (2006). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33(4):275–306.

- Plan, Y. and Vershynin, R. (2011). One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297.
- Ram, P. and Gray, A. (2013). Which space partitioning tree to use for search? In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 21*, pages 656–664. MIT Press, Cambridge, MA.
- Rebbapragada, U. and Brodley, C. (2007). Class noise mitigation through instance weighting. In *Proceedings of the 18th European Conference on Machine Learning*, pages 708–715, Warsaw, Poland.
- Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.
- Shakhnarovich, G., Darrell, T., and Indyk, P. (2006). *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. MIT Press, Cambridge, MA.
- Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge.
- Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5:595–645.
- Tarlow, D., Swersky, K., Swersky, K., Charlin, L., Sutskever, I., and Zemel, R. (2013). Stochastic k -neighborhood selection for supervised and unsupervised learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 199–207, Atlanta, GA.
- Tenenbaum, J., Silva, V. D., and Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.
- Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.
- Wagner, T. (1971). Convergence of the nearest neighbor rule. *IEEE Transactions on Information Theory*, 17(5):566–571.

- Weber, R., Schek, H., and Blott, S. (1998). A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the 24th International Conference on Very Large Data Bases*, pages 194–205, New York, NY.
- Xu, L., Crammer, K., and Schuurmans, D. (2006). Robust support vector machine training via convex outlier ablation. In *Proceedings of the 21st Conference on Artificial Intelligence*, pages 536–542, Boston, MA.